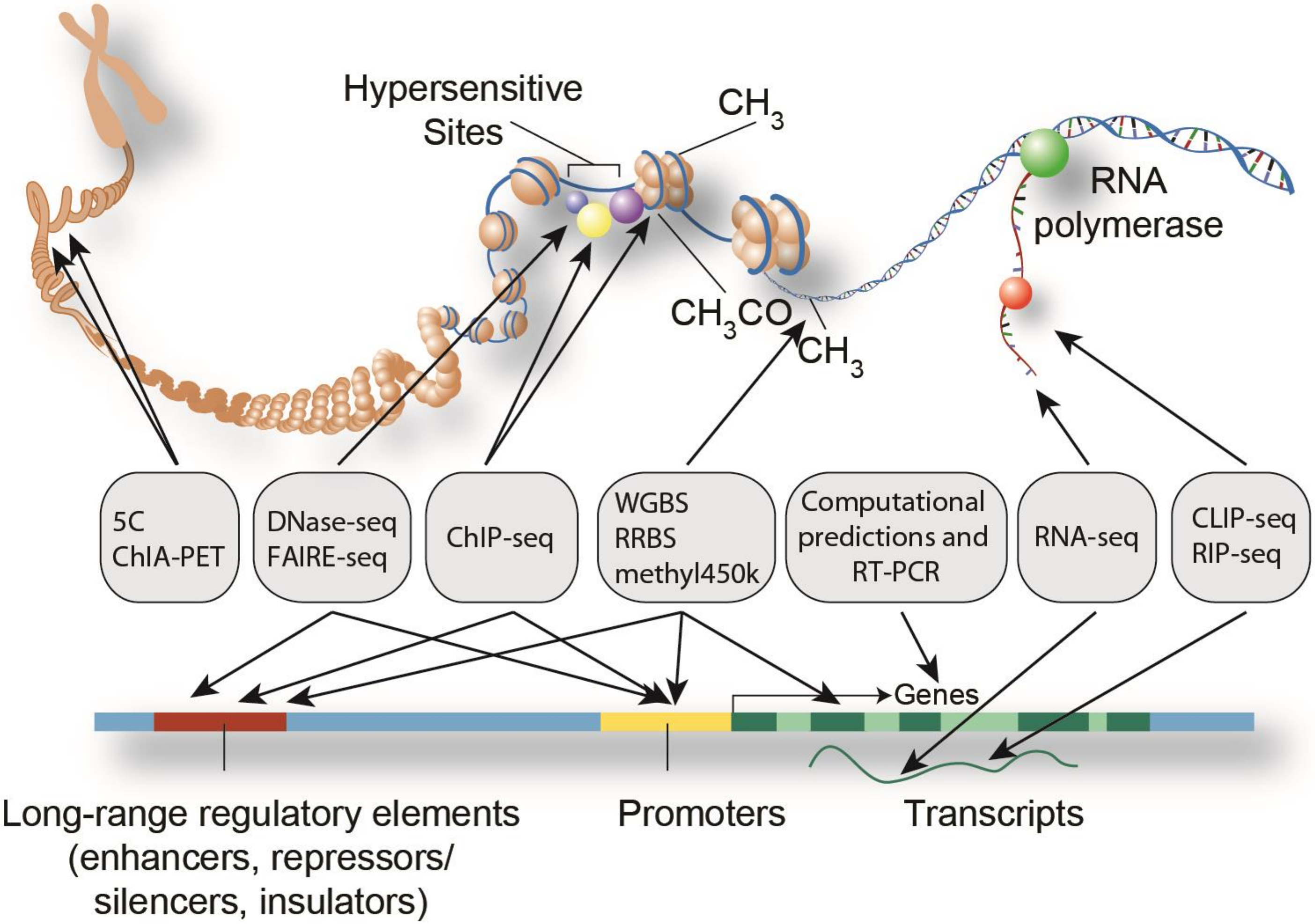
Summary of ENCODE Accomplishments

Michael Snyder
On Behalf of the ENCODE
Consortium

March 10, 2015

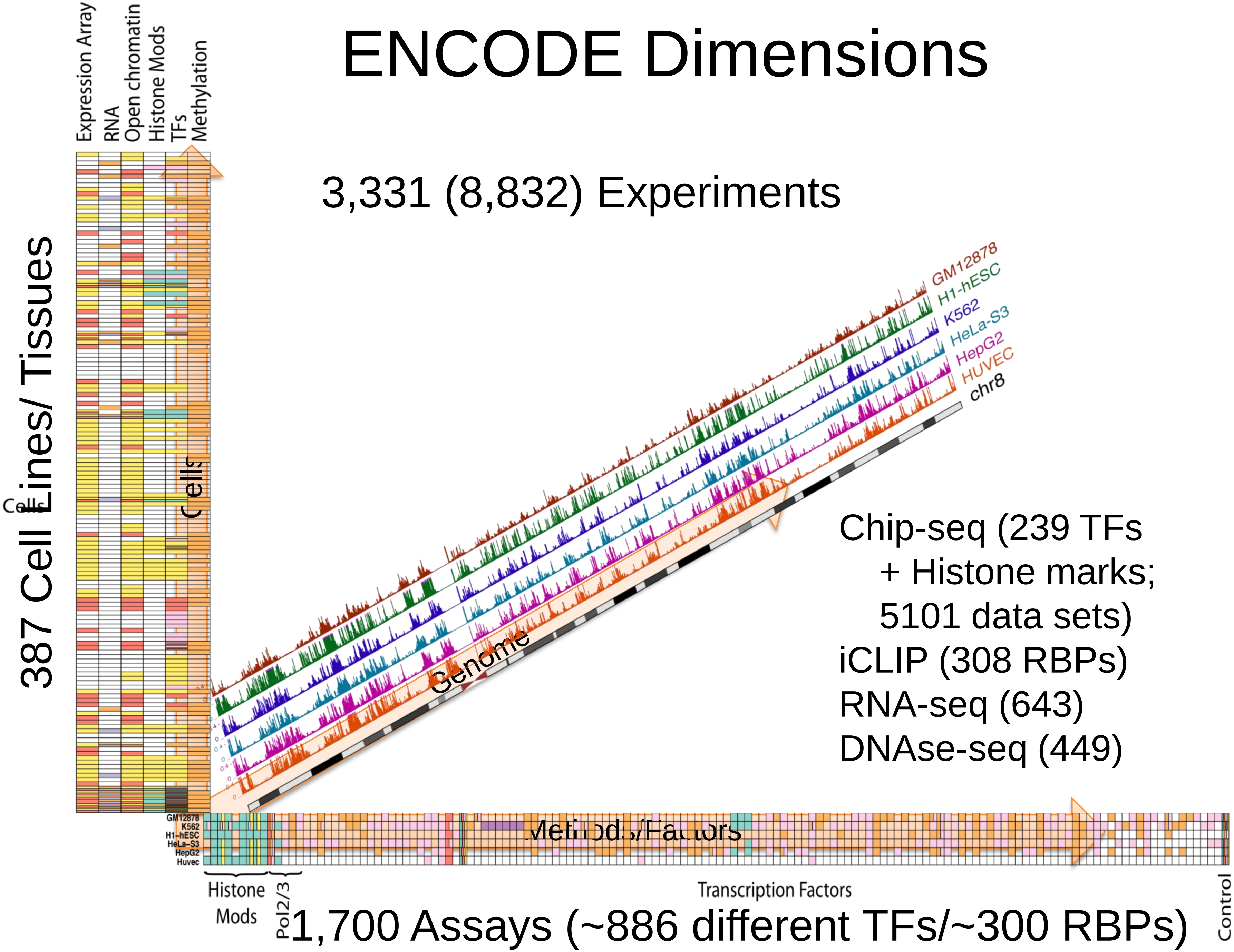
Conflicts: Personalis, Genapsys, AxioMx

ENCODE: Three Phases (2003-present)
Many Experimental Assays



Modified from PLoS Biol 9:e1001046, 2011
Science 306:636, 2004

ENCODE Dimensions



3,331 (8,832) Experiments

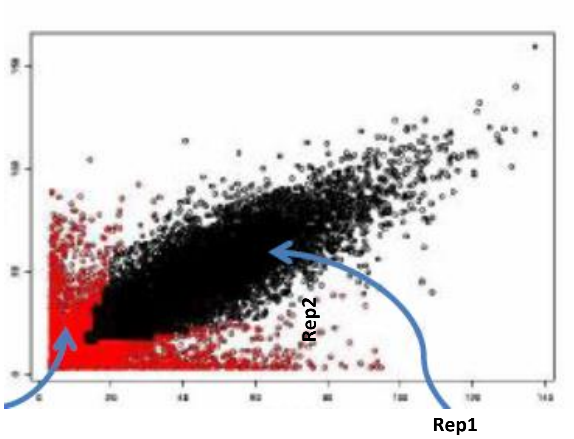
387 Cell Lines/ Tissues

Chip-seq (239 TFs
+ Histone marks;
5101 data sets)
iCLIP (308 RBPs)
RNA-seq (643)
DNAse-seq (449)

1,700 Assays (~886 different TFs/~300 RBPs)

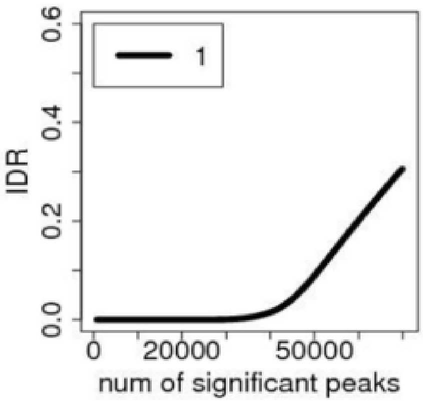
High Quality Data

- ≥ Two biological replicates
- Multiple quality control measures

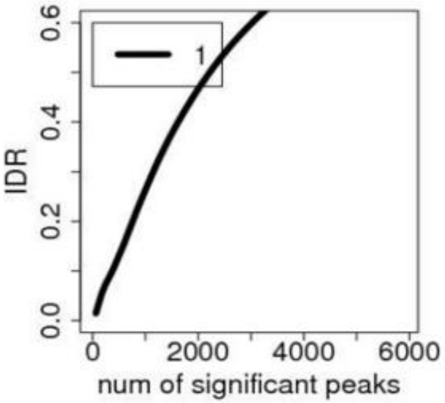


IDR Processing, QC and Blacklist Filtering

Good reproducibility

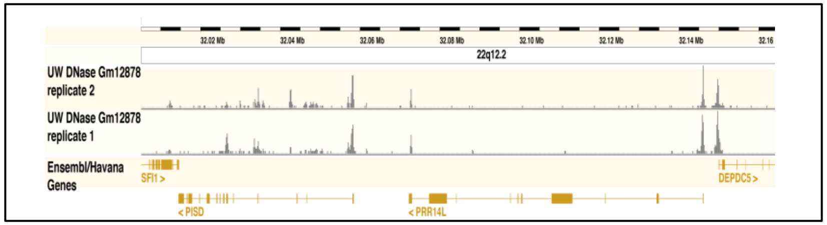


Poor reproducibility

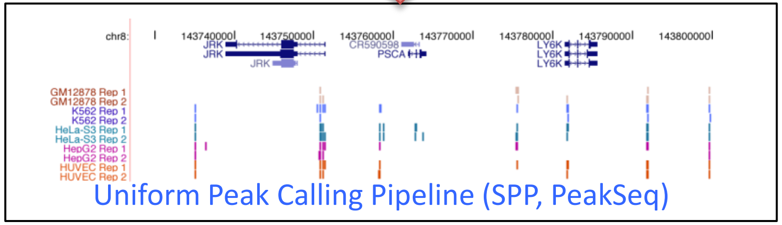


ENCODE Uniform Analysis Pipeline

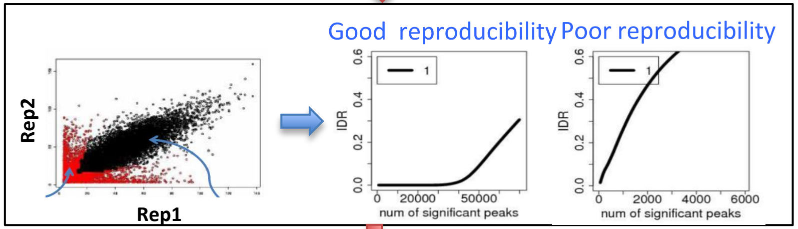
Anshul Kundaje



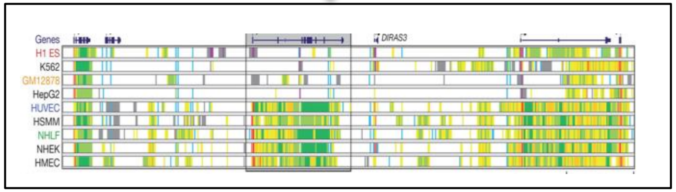
Mapped reads



Uniform peak Calling (SPP, PeakSeq)



Quality Control



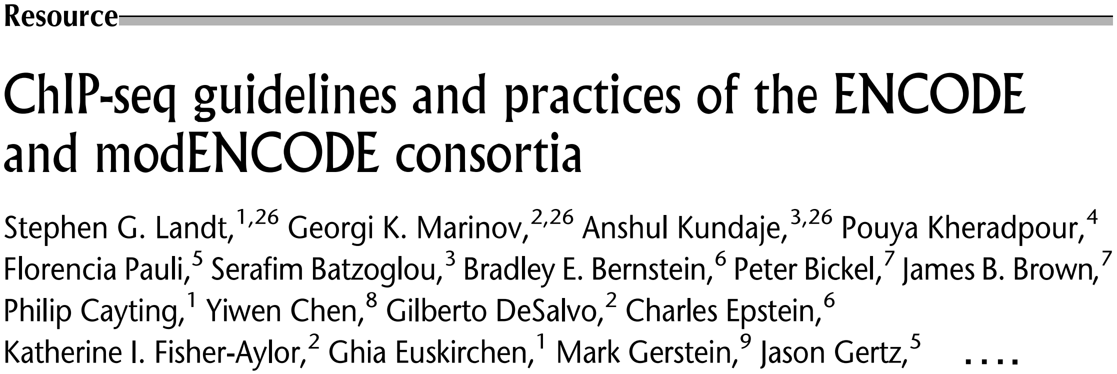
Derived Data (Chromosome Segments, Expression)

Processing & Element Calling Compatible with Other Projects: GTEx, REMC, IHEC

Established Standards For Community

- ChIP-Seq
- DNAseHS
- RNA-Seq

Antibody characterization, Biological replicates, QC measures



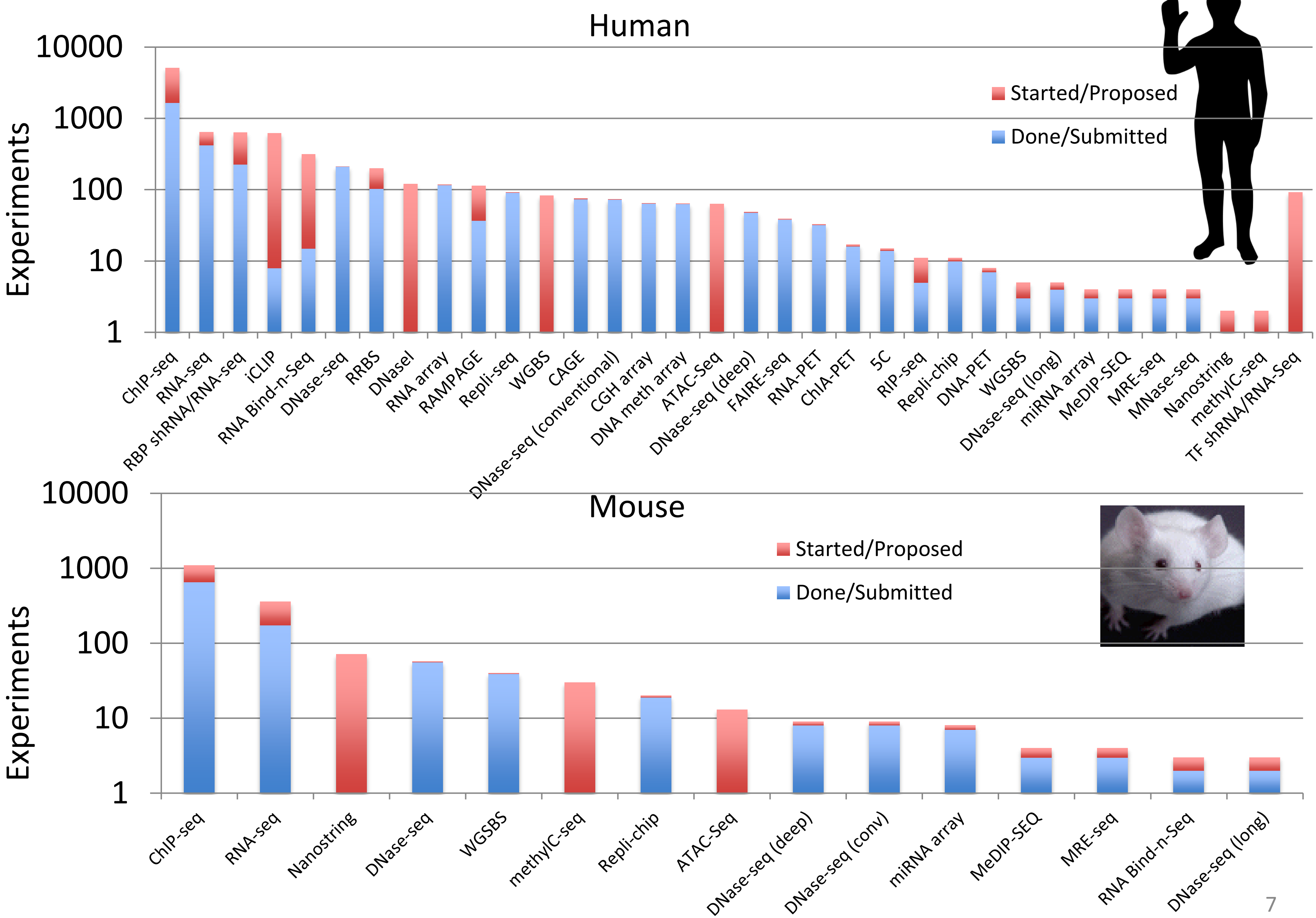
Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

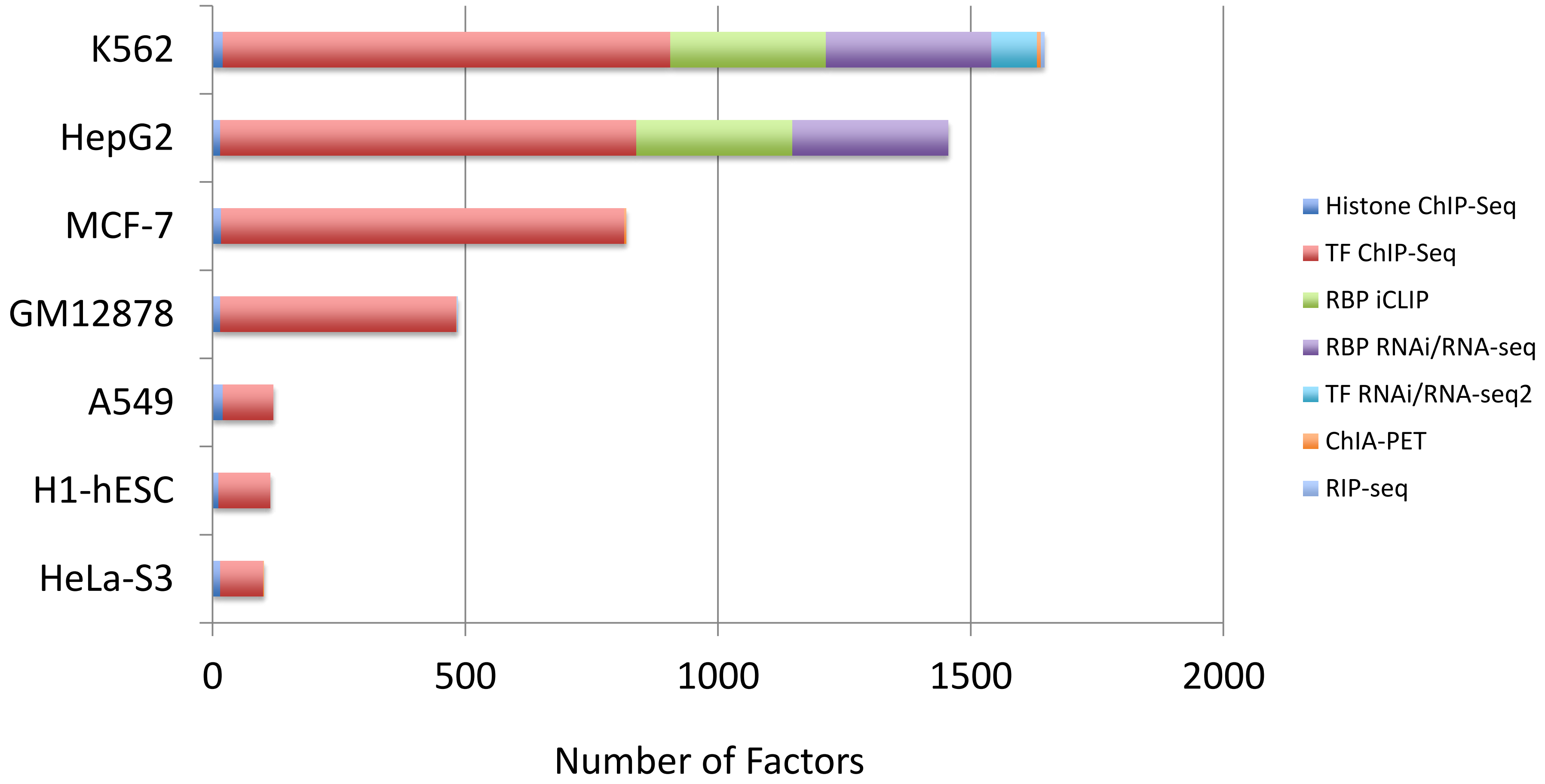
Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4]
Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7]
Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6]
Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] ….

Genome Res. 2012

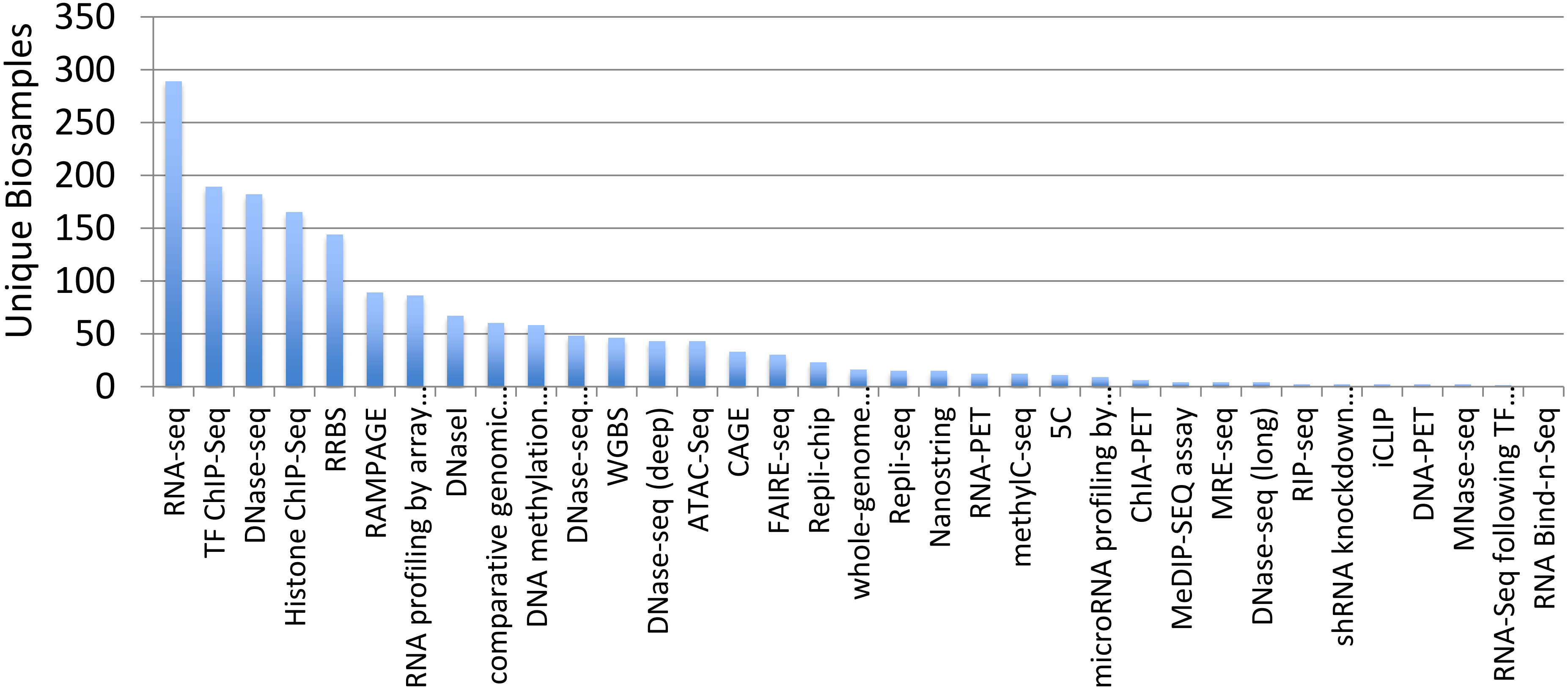
Assays/Data Types



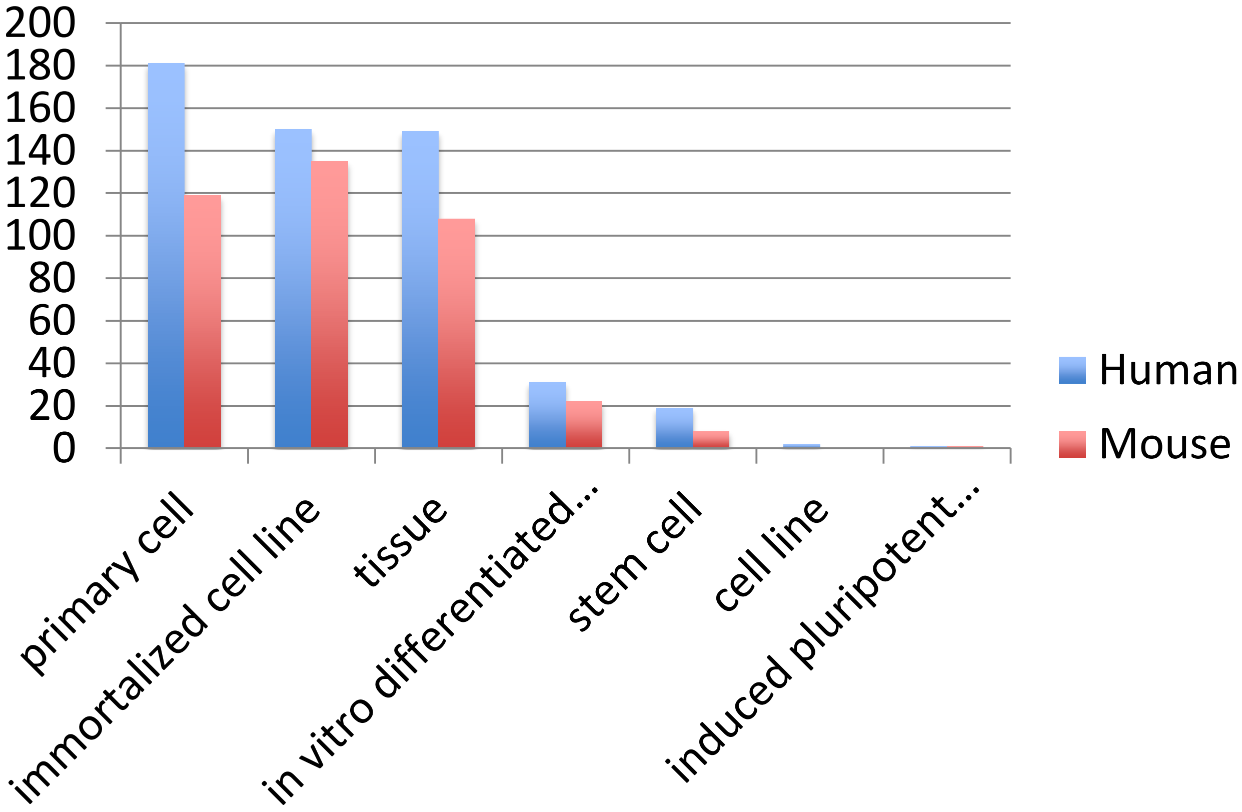
Deep Exploration of Some Lines Using Many Assays



Some Assays Were Conducted Across a Broad Range of Biosamples



Unique Biosample Types



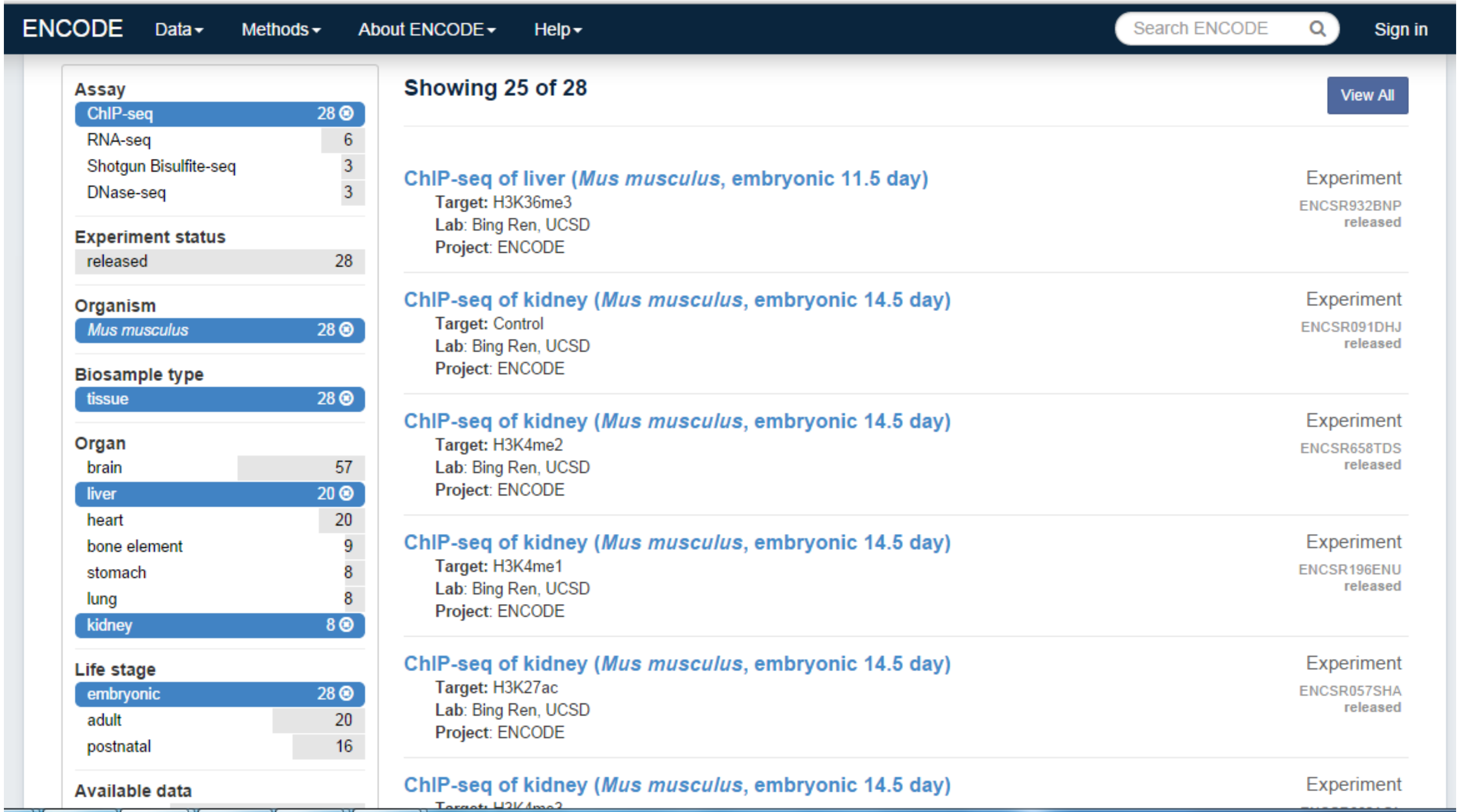
ENCODE Data

Cloud Storage and Computing

Data available at Amazon Web Services (AWS)

Uniform processing pipelines will be available at DNAnexus related projects

Highly Searchable



Assay

ChIP-seq	28
RNA-seq	6
Shotgun Bisulfite-seq	3
DNase-seq	3

Experiment status

released	28

Organism

Mus musculus	28

Biosample type

tissue	28

Organ

brain	57
liver	20
heart	20
bone element	9
stomach	8
lung	8
kidney	8

Life stage

embryonic	28
adult	20
postnatal	16

Available data

Showing 25 of 28

ChIP-seq of liver (*Mus musculus*, embryonic 11.5 day) — Experiment
Target: H3K36me3
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR932BNP released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) — Experiment
Target: Control
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR091DHJ released

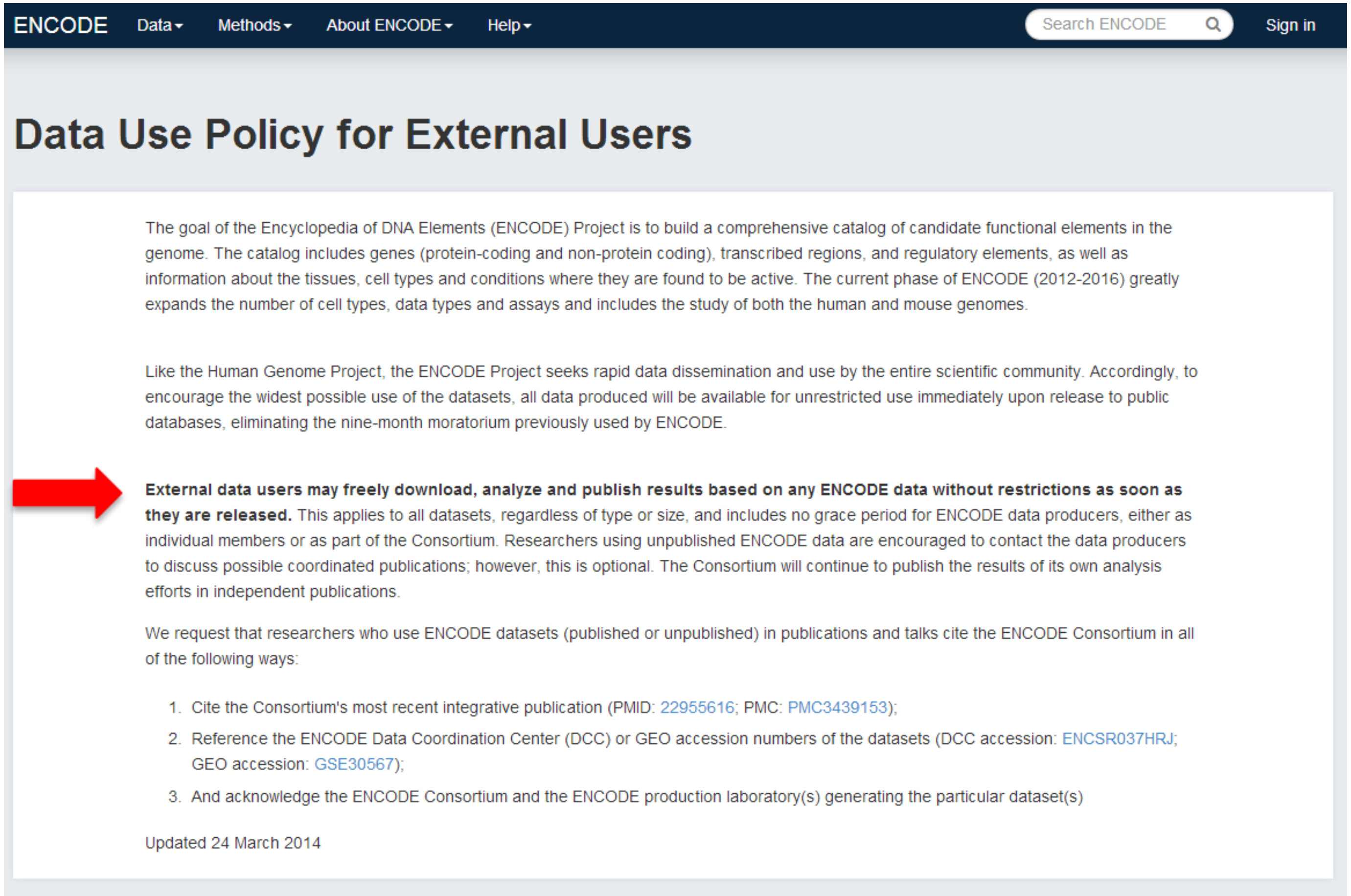
ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) — Experiment
Target: H3K4me2
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR658TDS released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) — Experiment
Target: H3K4me1
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR196ENU released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) — Experiment
Target: H3K27ac
Lab: Bing Ren, UCSD
Project: ENCODE
ENCSR057SHA released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day) — Experiment

ENCODE Data Open Access



Data Use Policy for External Users

The goal of the Encyclopedia of DNA Elements (ENCODE) Project is to build a comprehensive catalog of candidate functional elements in the genome. The catalog includes genes (protein-coding and non-protein coding), transcribed regions, and regulatory elements, as well as information about the tissues, cell types and conditions where they are found to be active. The current phase of ENCODE (2012-2016) greatly expands the number of cell types, data types and assays and includes the study of both the human and mouse genomes.

Like the Human Genome Project, the ENCODE Project seeks rapid data dissemination and use by the entire scientific community. Accordingly, to encourage the widest possible use of the datasets, all data produced will be available for unrestricted use immediately upon release to public databases, eliminating the nine-month moratorium previously used by ENCODE.

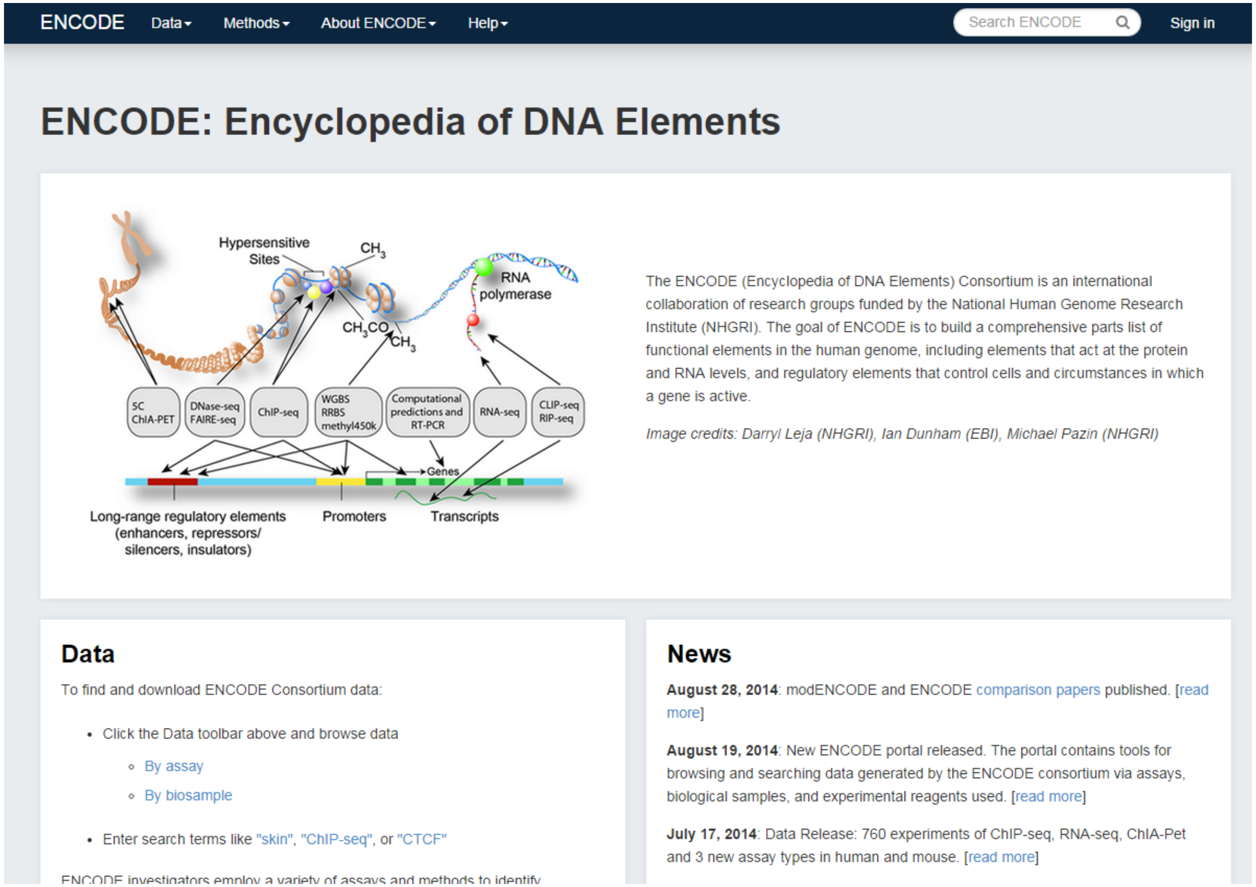
External data users may freely download, analyze and publish results based on any ENCODE data without restrictions as soon as they are released. This applies to all datasets, regardless of type or size, and includes no grace period for ENCODE data producers, either as individual members or as part of the Consortium. Researchers using unpublished ENCODE data are encouraged to contact the data producers to discuss possible coordinated publications; however, this is optional. The Consortium will continue to publish the results of its own analysis efforts in independent publications.

We request that researchers who use ENCODE datasets (published or unpublished) in publications and talks cite the ENCODE Consortium in all of the following ways:

1. Cite the Consortium's most recent integrative publication (PMID: 22955616; PMC: PMC3439153);
2. Reference the ENCODE Data Coordination Center (DCC) or GEO accession numbers of the datasets (DCC accession: ENCSR037HRJ; GEO accession: GSE30567);
3. And acknowledge the ENCODE Consortium and the ENCODE production laboratory(s) generating the particular dataset(s)

Updated 24 March 2014

New ENCODE Portal
https://www.encodeproject.org



ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Data

To find and download ENCODE Consortium data:

- Click the Data toolbar above and browse data
  - By assay
  - By biosample
- Enter search terms like "skin", "ChIP-seq", or "CTCF"

ENCODE investigators employ a variety of assays and methods to identify

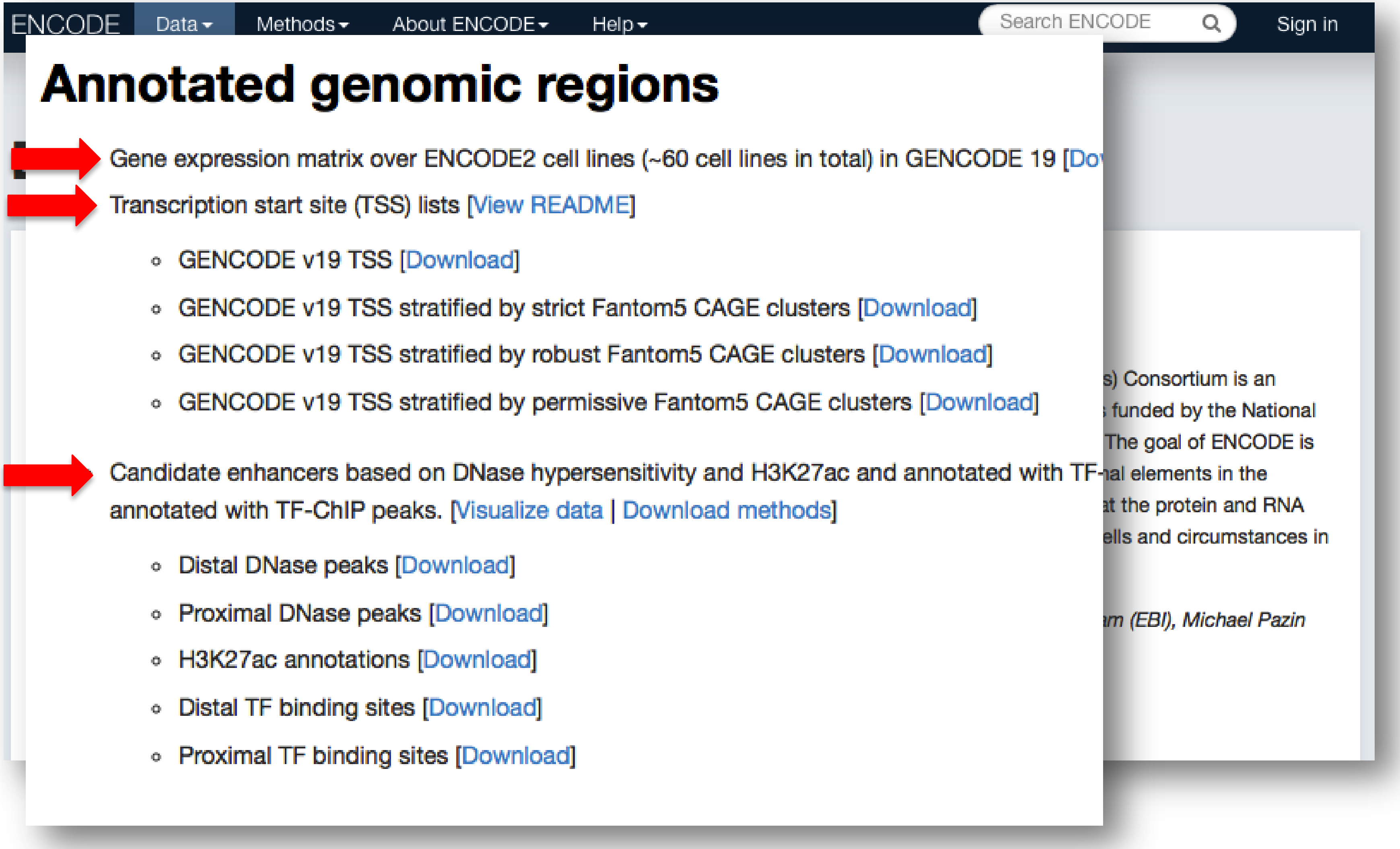
News

August 28, 2014: modENCODE and ENCODE comparison papers published. [read more]

August 19, 2014: New ENCODE portal released. The portal contains tools for browsing and searching data generated by the ENCODE consortium via assays, biological samples, and experimental reagents used. [read more]

July 17, 2014: Data Release: 760 experiments of ChIP-seq, RNA-seq, ChIA-Pet and 3 new assay types in human and mouse. [read more]

ENCODE Encyclopedia Prototype



Annotated genomic regions

Gene expression matrix over ENCODE2 cell lines (~60 cell lines in total) in GENCODE 19 [Do

Transcription start site (TSS) lists [View README]

- GENCODE v19 TSS [Download]
- GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [Download]
- GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [Download]
- GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [Download]

Candidate enhancers based on DNase hypersensitivity and H3K27ac and annotated with TF-annotated with TF-ChIP peaks. [Visualize data | Download methods]

- Distal DNase peaks [Download]
- Proximal DNase peaks [Download]
- H3K27ac annotations [Download]
- Distal TF binding sites [Download]
- Proximal TF binding sites [Download]

s) Consortium is an
funded by the National
The goal of ENCODE is
al elements in the
at the protein and RNA
ells and circumstances in

am (EBI), Michael Pazin

http://encodeproject.org

10 Computational Groups

Analyzing data in a variety of different ways

- GWAS
- Cancer
- Regulatory principles

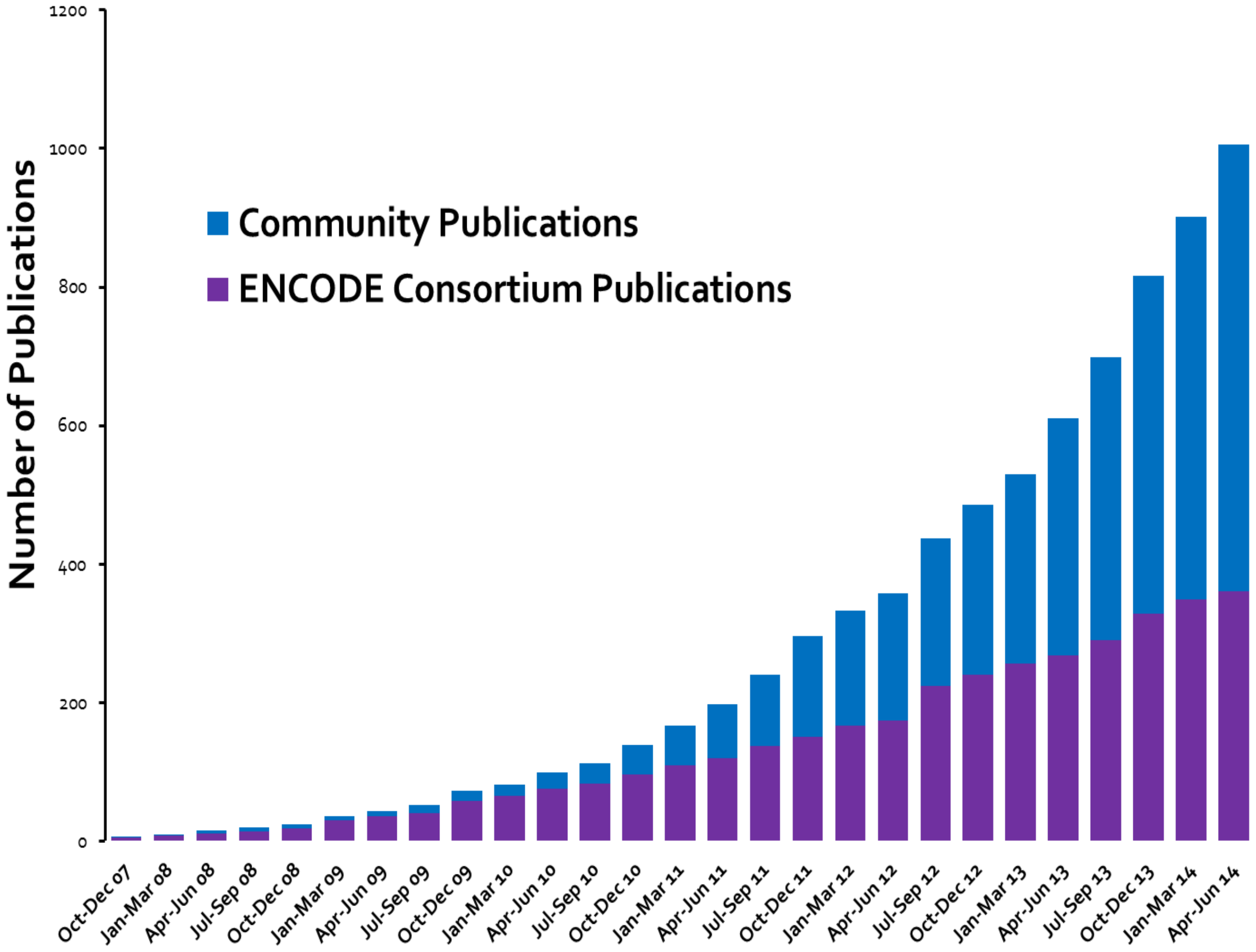
Software Tools

- >30 Different algorithms
- Wide variety of areas. Examples:
  - Segmentation
  - Allele calling
  - 3D nuclear analysis
  - Data processing and peak calling
  - Data quality control

Summary of Impact

- Lots of open diverse data types on same cell lines/tissues
- Experimental standards
- New analysis methods
- Methods and standards adopted by other large communities: e.g GTEx, REMC, betaCell, CIRM
- Data are widely used

ENCODE Publications

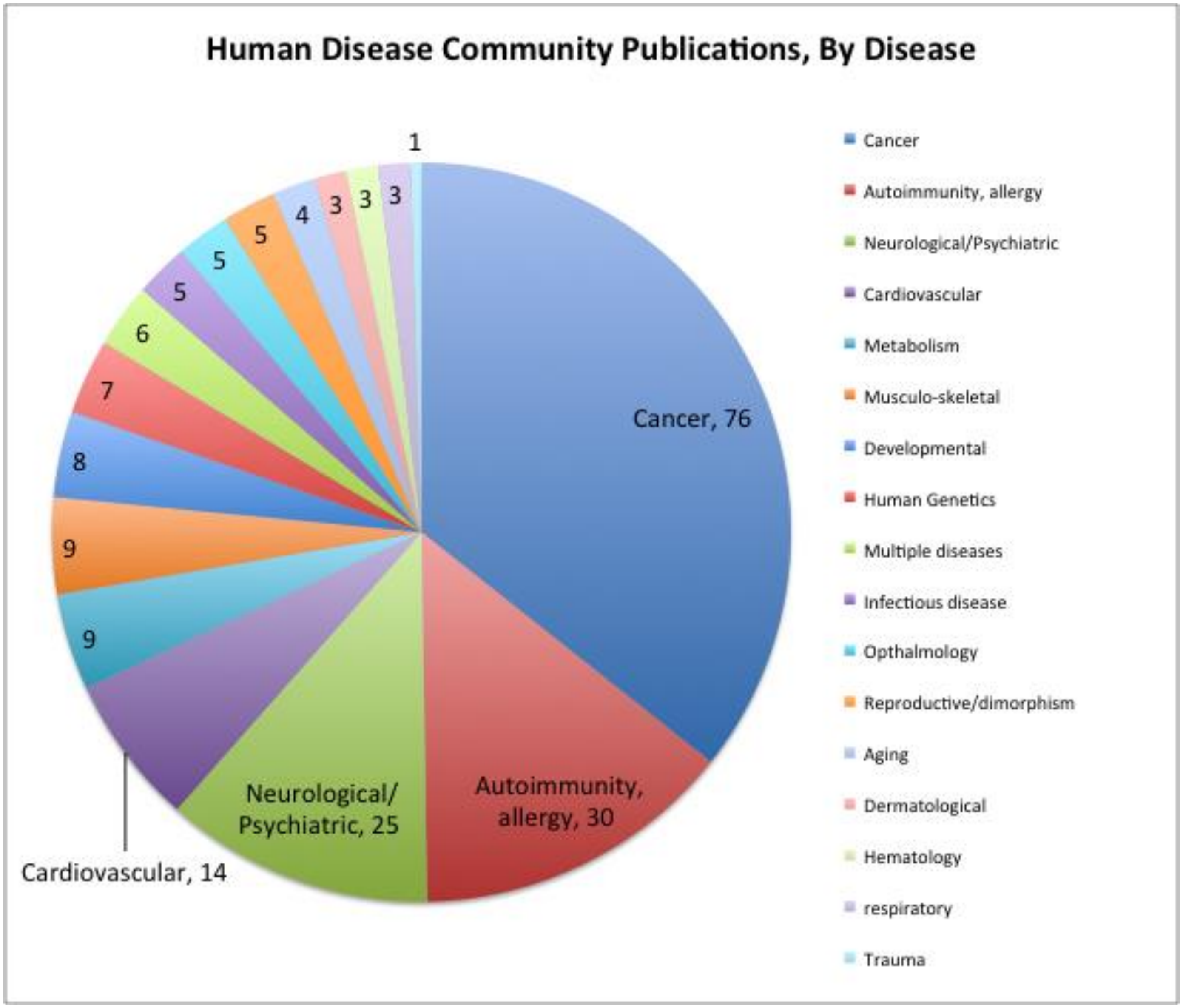


Feb 2015

>750 Papers NonENCODE

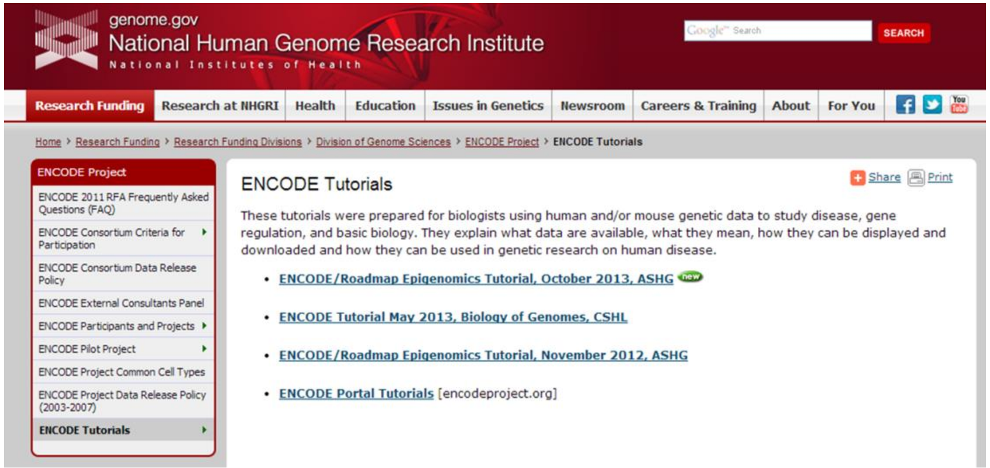
+ 150 mod-ENCODE

ENCODE Community Publications



Outreach Activities

- Tutorials:https://www.encodeproject.org/tutorials
- (http://www.genome.gov/27553900)


genome.gov
National Human Genome Research Institute
National Institutes of Health
Google Search
SEARCH
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
For You
Home > Research Funding > Research Funding Divisions > Division of Genome Sciences > ENCODE Project > ENCODE Tutorials
ENCODE Project
ENCODE 2011 RFA Frequently Asked Questions (FAQ)
ENCODE Consortium Criteria for Participation
ENCODE Consortium Data Release Policy
ENCODE External Consultants Panel
ENCODE Participants and Projects
ENCODE Pilot Project
ENCODE Project Common Cell Types
ENCODE Project Data Release Policy (2003-2007)
ENCODE Tutorials
Share
Print
new

ENCODE Tutorials

These tutorials were prepared for biologists using human and/or mouse genetic data to study disease, gene regulation, and basic biology. They explain what data are available, what they mean, how they can be displayed and downloaded and how they can be used in genetic research on human disease.

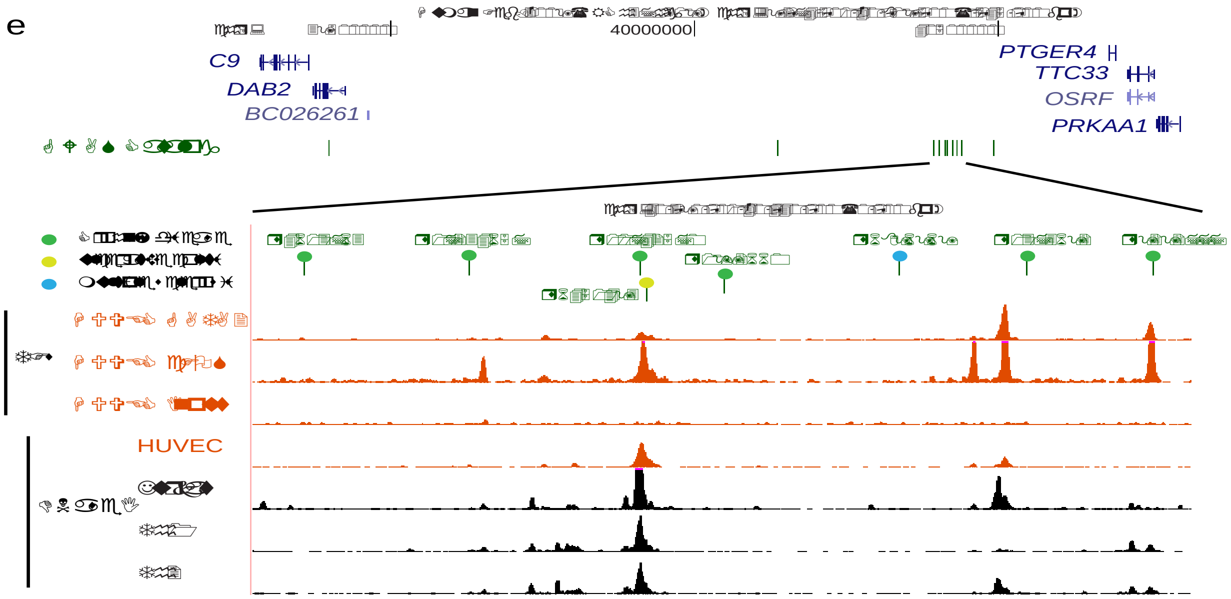
- ENCODE/Roadmap Epigenomics Tutorial, October 2013, ASHG
- ENCODE Tutorial May 2013, Biology of Genomes, CSHL
- ENCODE/Roadmap Epigenomics Tutorial, November 2012, ASHG
- ENCODE Portal Tutorials [encodeproject.org]

- CHARGE-ENCODE workshop
- User’s meeting in 2015

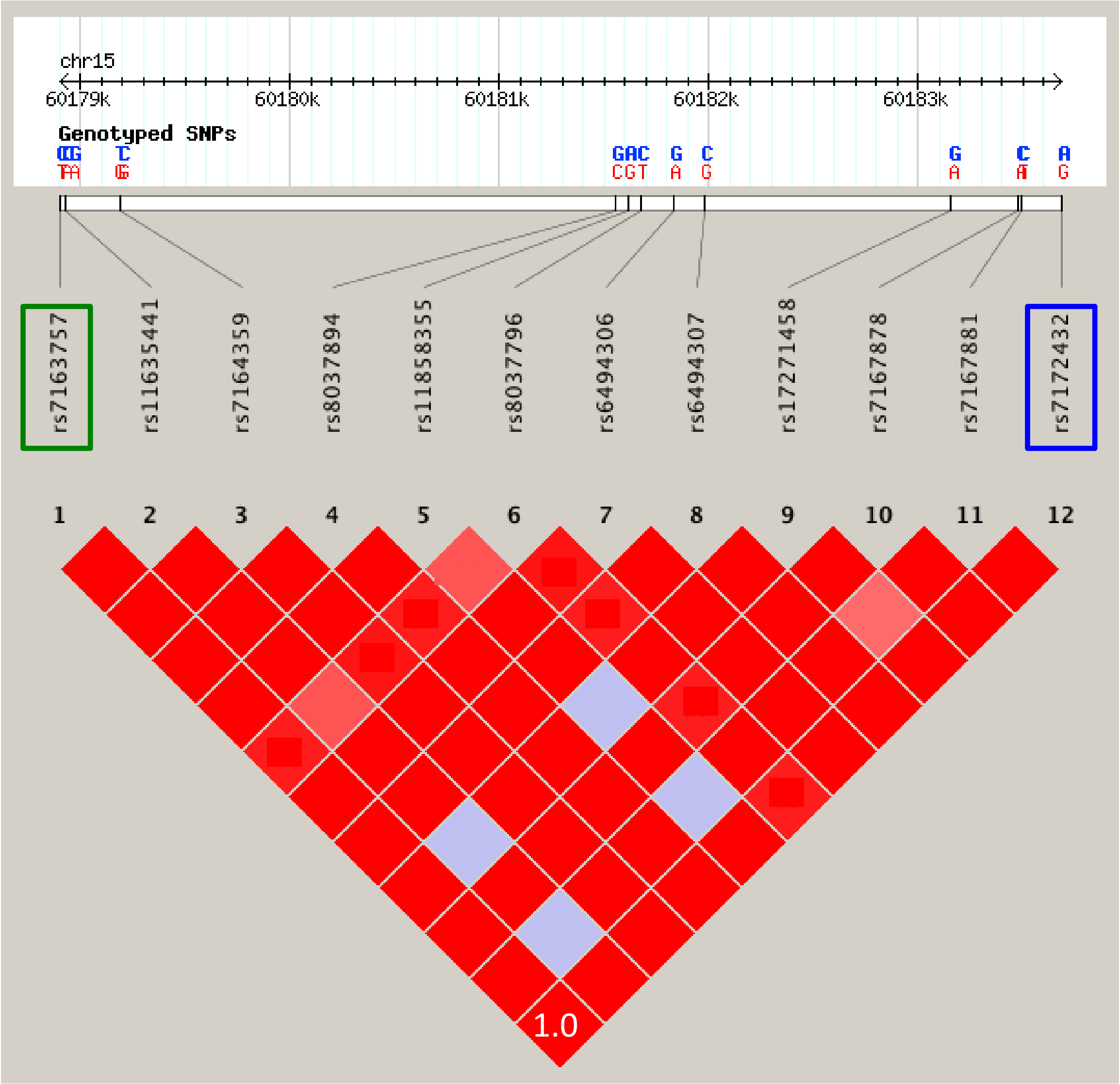
Additional High Level Impact

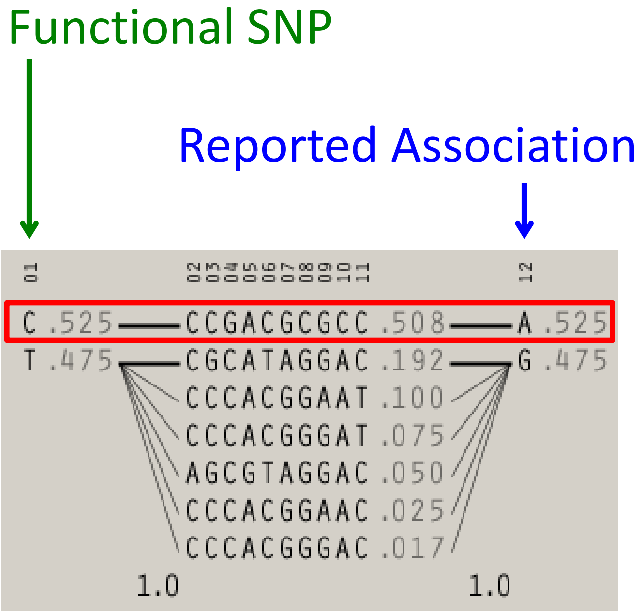
1) Segmenting genome into types of elements
2) Gene regulatory principles
3) GWAS

>85% of lead SNPs lie outside of coding regions



Example: rs7172432 in Type 2 Diabetes





Functional SNP

Reported Association

GWAS (Japanese): T Yamauchi *et al. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B***.** Nature Genetics 42, 864–868 (2010).

GWAS (Danish): N. Grarup *et al. The diabetogenic VPS13C/C2CD4A/C2CD4B rs7172432 variant impairs glucose-stimulated insulin response in 5,722 non-diabetic Danish individuals.* Diabetologia (2011) 54:789–794

Altered View of the Human Genome

2003

- 25,000 Protein Coding Genes (1.5%)
- Few Non Coding Genes (Mostly tRNAs, snoRNAs)
- Little regulatory information mapped

2015

- 20,000 Protein Coding Genes
- Thousands of noncoding genes
- More potential regulatory DNA than protein coding DNA

The ENCODE 3 Consortium

http://www.genome.gov/26525220

Alternative/Additional Slides

Three Phases

I) Pilot Phase -1% of Genome (2003-2007)

II) Scale Up Phase I (2007-2012)

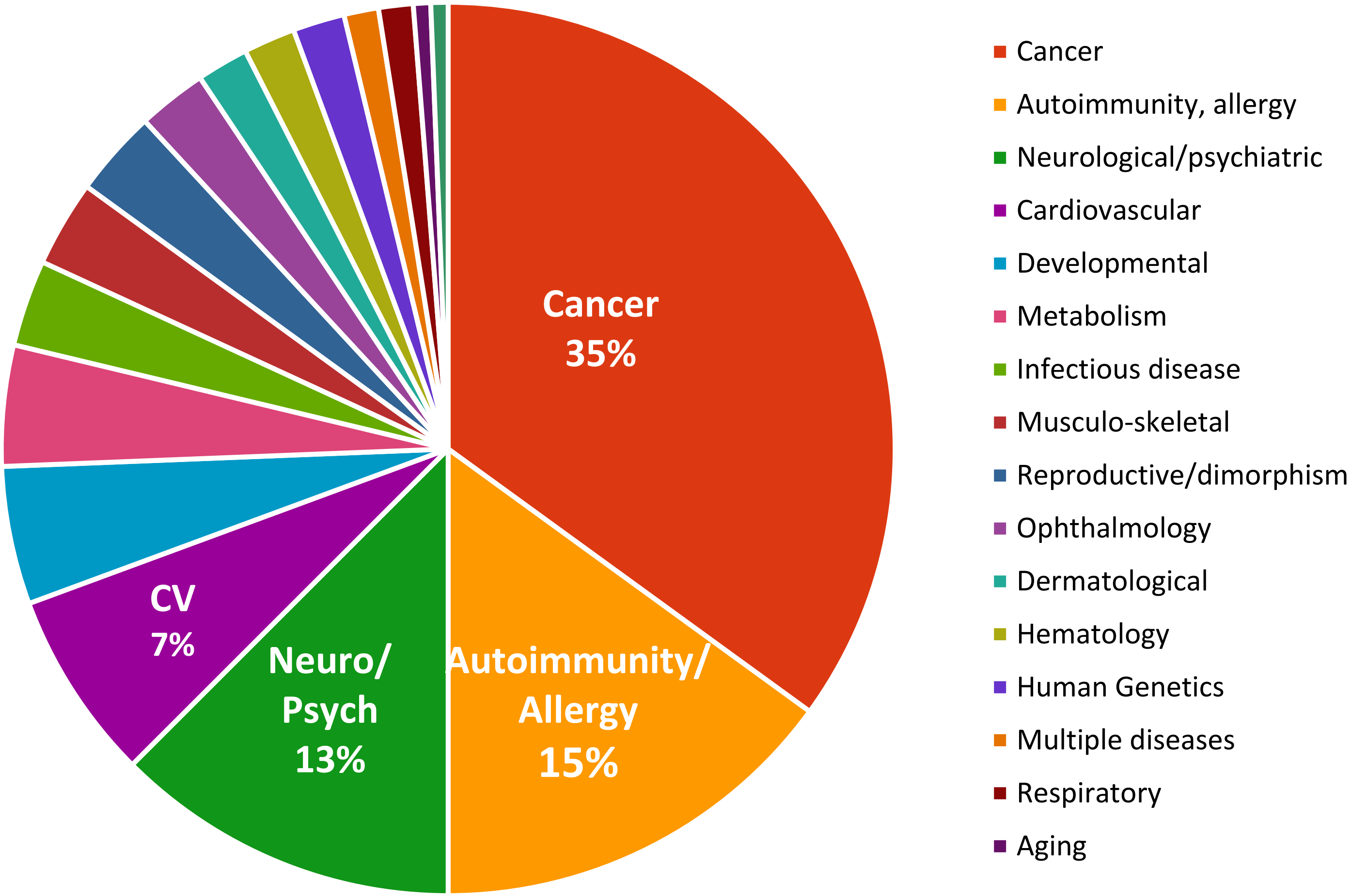
III) Current Production Phase (2012-2016)

Related Projects:

Mouse ENCODE (2009-2012)

modENCODE (2007-2012)

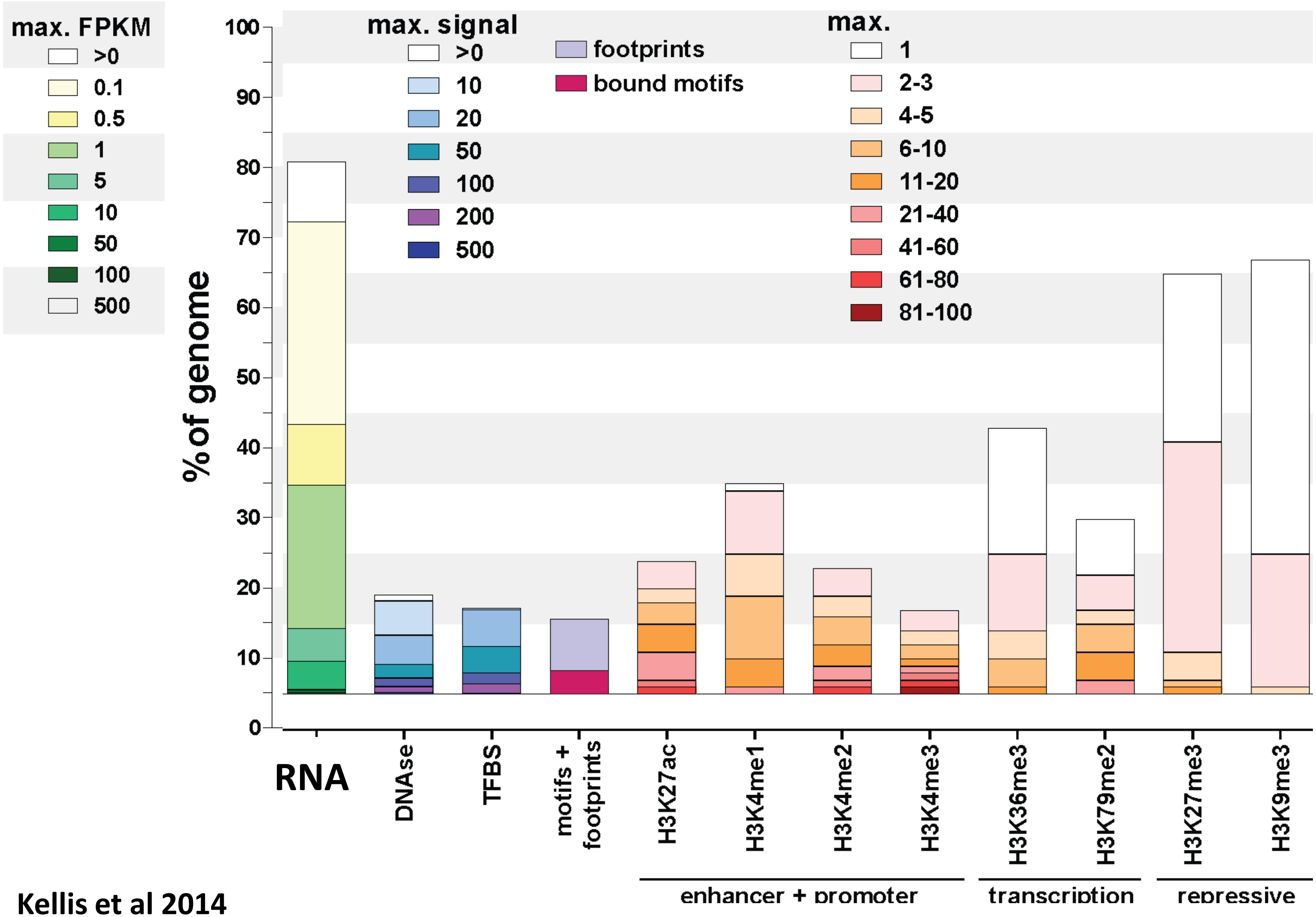
Categories of Disease-Related ENCODE Community Publications



Standard ENCODE Use Cases: Hypothesis Generation

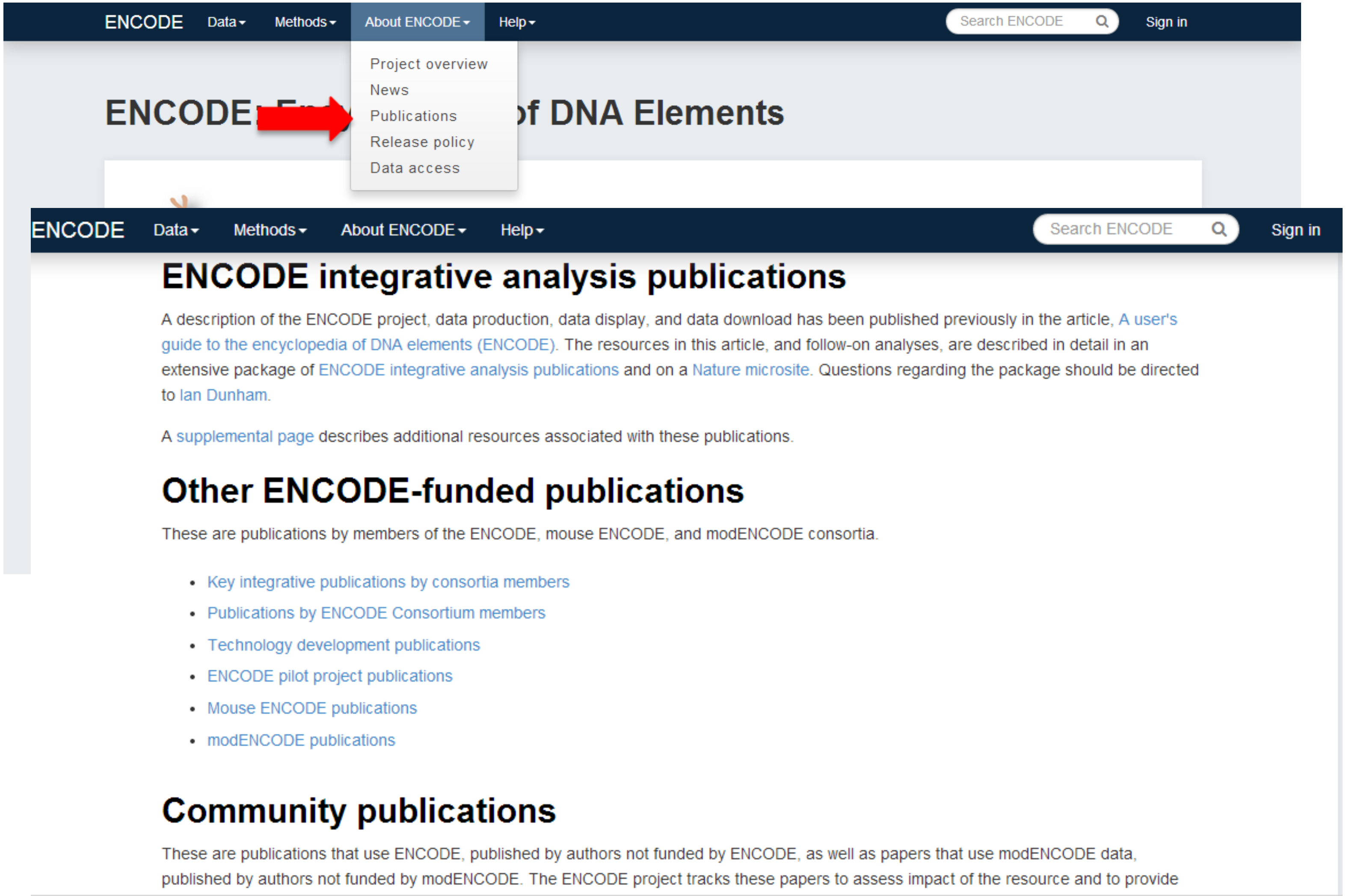
- Prediction of:
  - causal variants/regulatory elements
  - target genes
  - target cell types
  - mechanism for phenotype changes

The Genome Is Active!



Kellis et al 2014

ENCODE Publications



ENCODE integrative analysis publications

A description of the ENCODE project, data production, data display, and data download has been published previously in the article, A user's guide to the encyclopedia of DNA elements (ENCODE). The resources in this article, and follow-on analyses, are described in detail in an extensive package of ENCODE integrative analysis publications and on a Nature microsite. Questions regarding the package should be directed to Ian Dunham.

A supplemental page describes additional resources associated with these publications.

Other ENCODE-funded publications

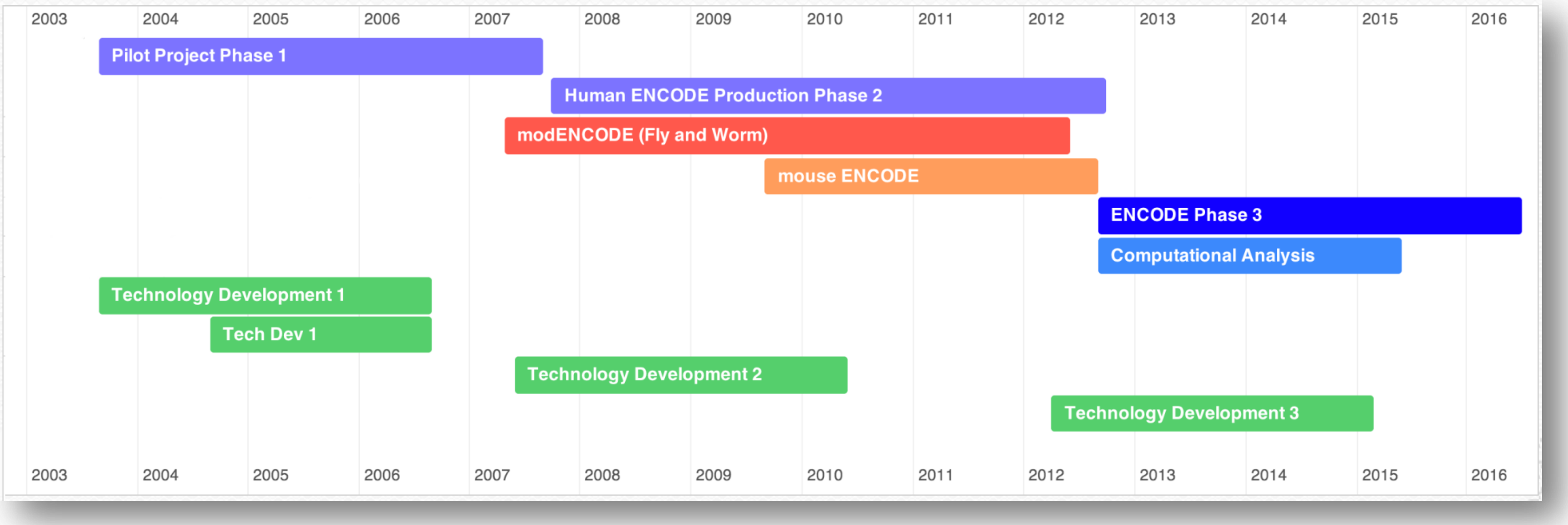
These are publications by members of the ENCODE, mouse ENCODE, and modENCODE consortia.

- Key integrative publications by consortia members
- Publications by ENCODE Consortium members
- Technology development publications
- ENCODE pilot project publications
- Mouse ENCODE publications
- modENCODE publications

Community publications

These are publications that use ENCODE, published by authors not funded by ENCODE, as well as papers that use modENCODE data, published by authors not funded by modENCODE. The ENCODE project tracks these papers to assess impact of the resource and to provide

ENCODE Timeline



The ENCODE Consortium Phase 3

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom) **Mike Cherry**

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

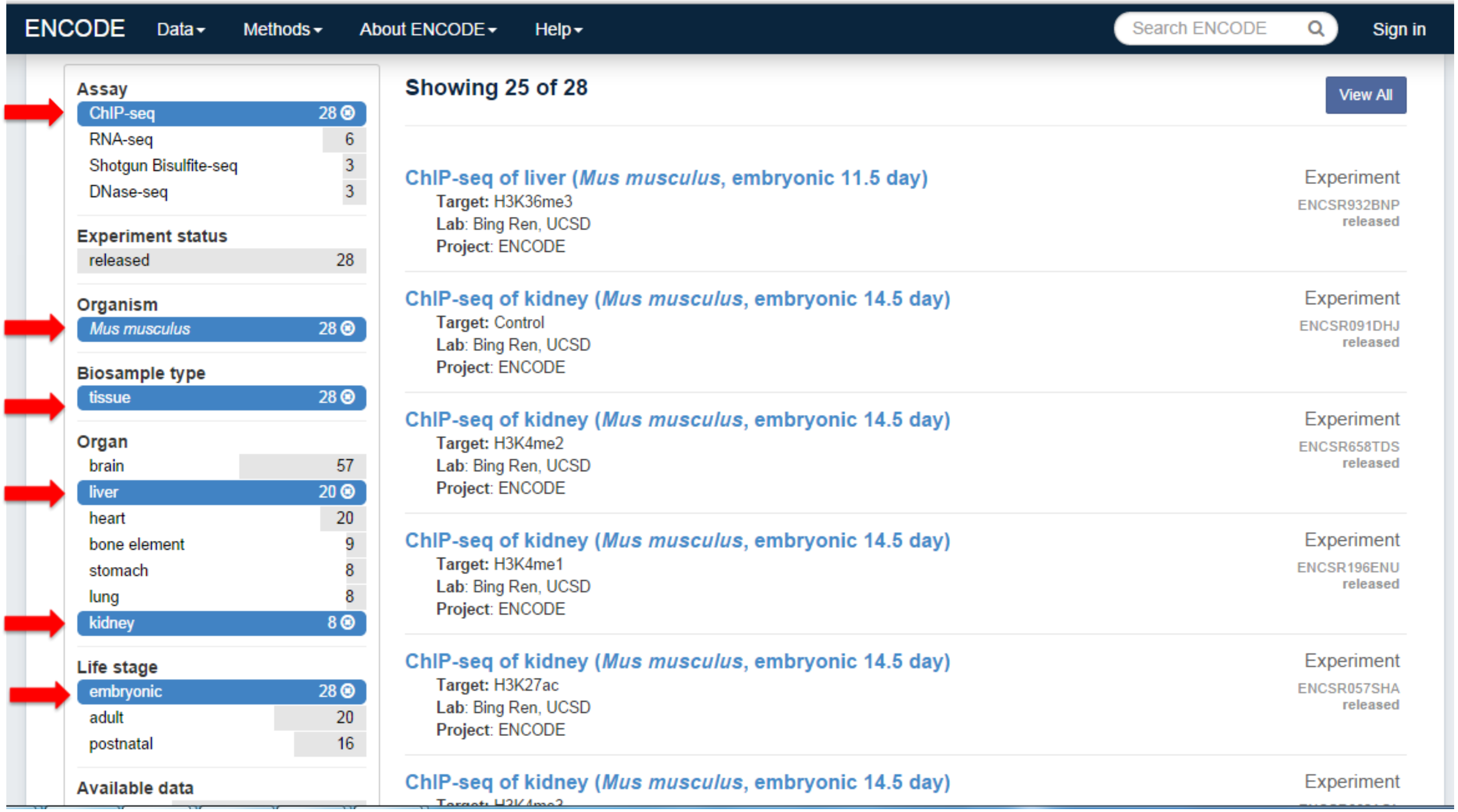
Zhiping Weng (Nathan Trinklein, Rick Myers)

Brenton Graveley (John Rinn, Others)

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good

Metadata-driven searches



Assay

ChIP-seq	28
RNA-seq	6
Shotgun Bisulfite-seq	3
DNase-seq	3

Experiment status

released	28

Organism

Mus musculus	28

Biosample type

tissue	28

Organ

brain	57
liver	20
heart	20
bone element	9
stomach	8
lung	8
kidney	8

Life stage

embryonic	28
adult	20
postnatal	16

Available data

Showing 25 of 28

ChIP-seq of liver (*Mus musculus*, embryonic 11.5 day)
Target: H3K36me3
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR932BNP
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: Control
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR091DHJ
released

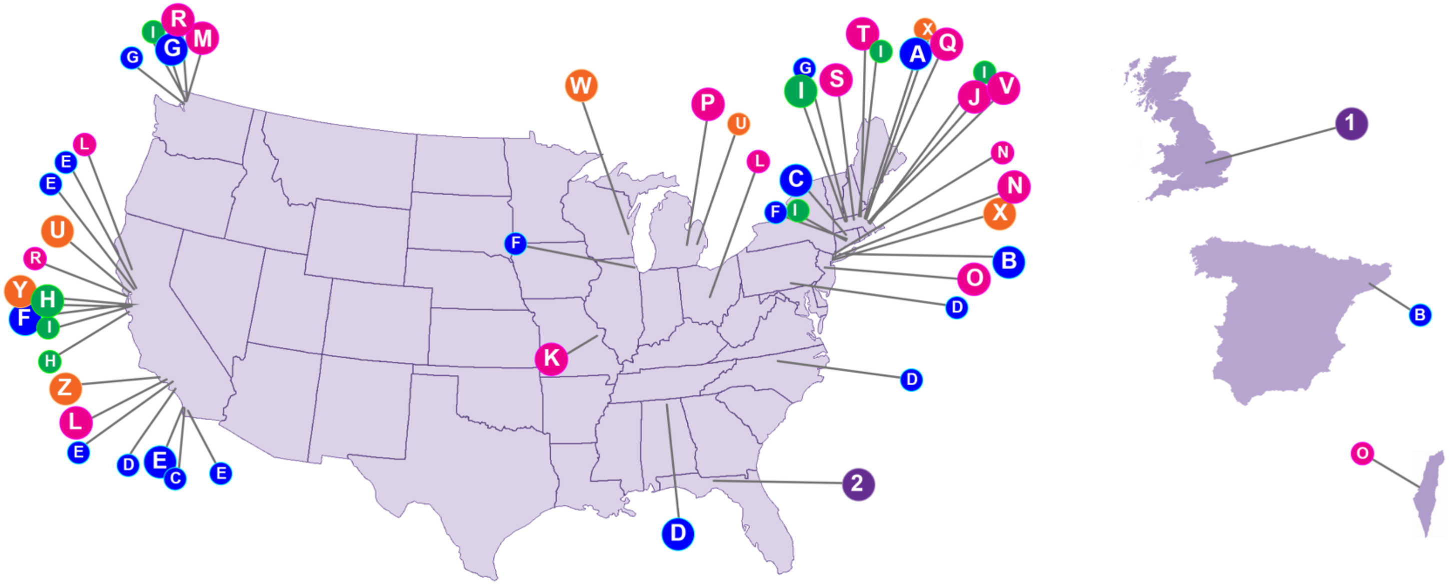
ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me2
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR658TDS
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me1
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR196ENU
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K27ac
Lab: Bing Ren, UCSD
Project: ENCODE
Experiment
ENCSR057SHA
released

ChIP-seq of kidney (*Mus musculus*, embryonic 14.5 day)
Target: H3K4me3
Experiment

ENCODE Consortium Phase 3



Production Groups

- A **Broad Institute**
- B **Cold Spring Harbor;** Centre for Genomic Regulation (CRG);
- C **University of Connecticut Health Center**; UCSD
- D **HudsonAlpha;** Pennsylvania State; UC Irvine; Duke; Caltech
- E **UCSD**; Salk Institute ; Joint Genome Institute; Lawrence Berkeley National Laboratory; UCSD
- F **Stanford**; University of Chicago; Yale
- G **University of Washington**; Fred Hutchinson Cancer Research Center; University of Massachusetts Medical School

Data Coordination Center

- H **Stanford**; UCSC

Data Analysis Center

- I **University of Massachusetts Medical School**; Yale; MIT; Stanford; Harvard; University of Washington

Technology Development Groups

- J **MIT**
- K **Washington University, St. Louis**
- L **USC**; Ohio State University; UC, Davis
- M **University of Washington**
- N **Sloan-Kettering**; Weill Cornell Medical College
- O **Princeton;** Weizmann
- P **University of Michigan**
- Q **Broad Institute**
- R **University of Washington**; UCSF
- S **Advanced RNA Technologies, LLC**
- T **Harvard**

Computational Analysis Groups

- U **Berkeley**; Wayne State University
- V **MIT**
- W **University of Wisconsin**
- X **Sloan-Kettering**; Broad Institute
- Y **Stanford**
- Z **UCLA**

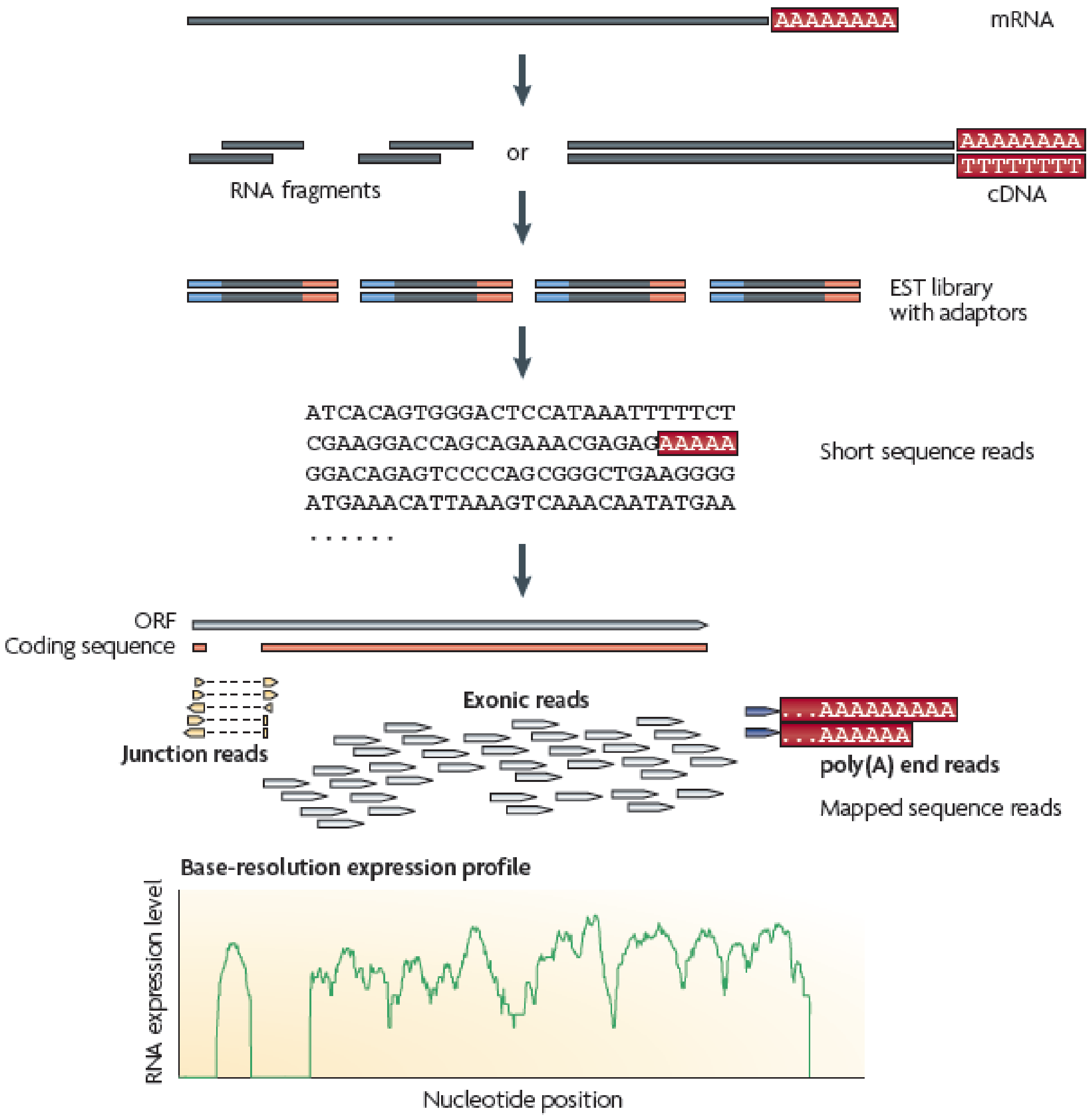
Affiliated Groups

- 1 **Wellcome Trust Sanger Institute**
- 2 **Florida State University**

Goals of ENCODE

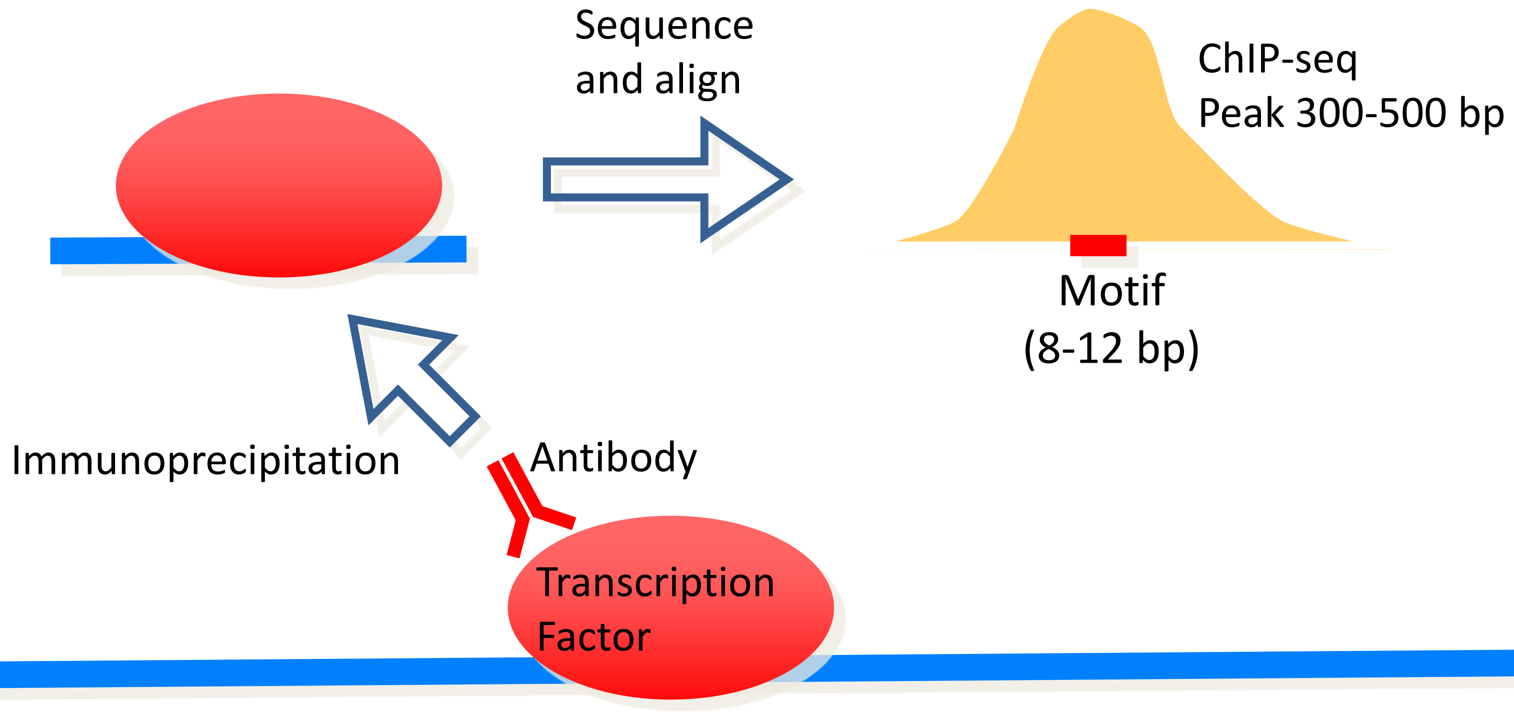
- Catalog the functional elements in human and mouse genomes
- Generate high quality data using high throughput pipelines
- Develop new technologies and analytical tools to generate, analyze and validate data
- Provide data and tools to the community in as useful form as possible

RNA-Sequencing



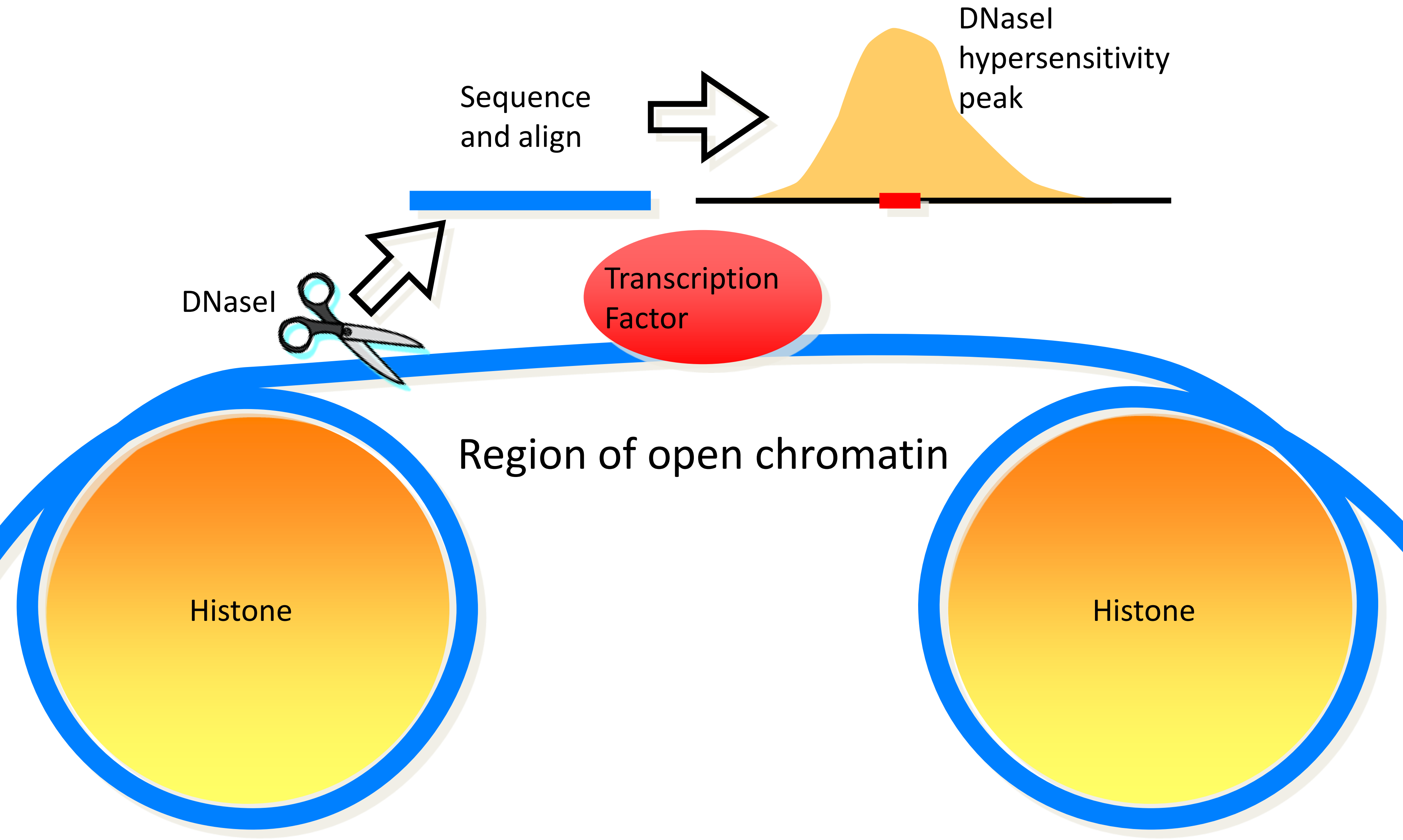
Wang et al. 2009 Nat Gen. Rev.

Functional data: ChIP-seq

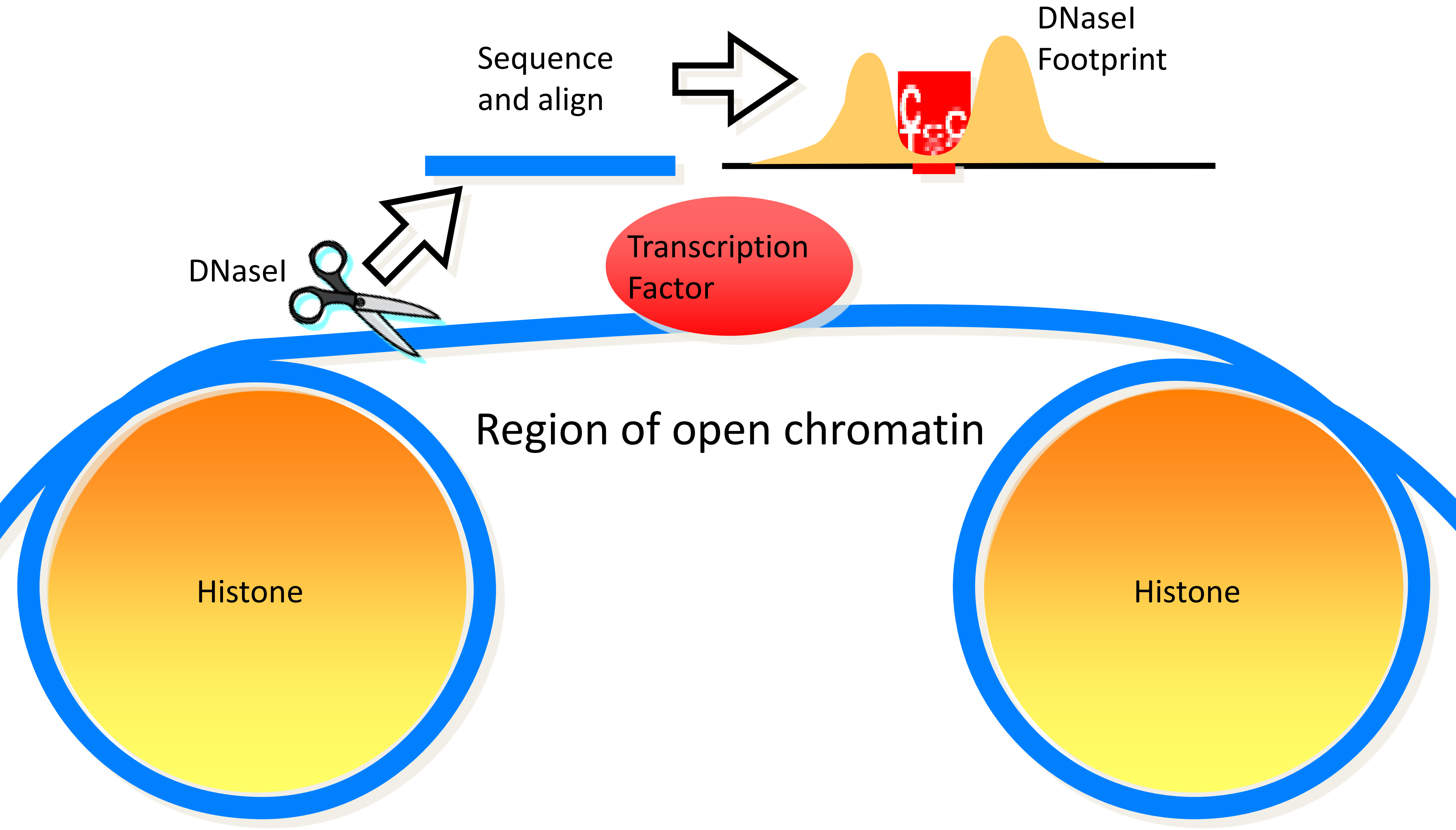


ChIP-exo
Histone Marks

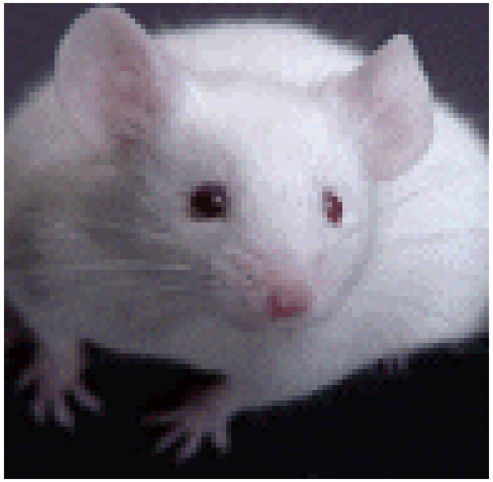
Functional data: DNase-seq

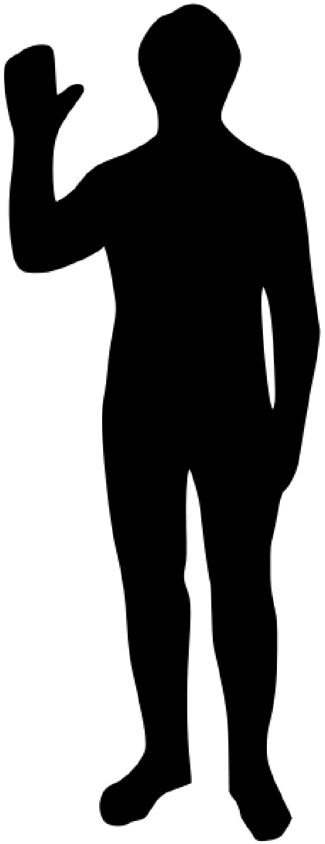


Functional data: DNase footprints

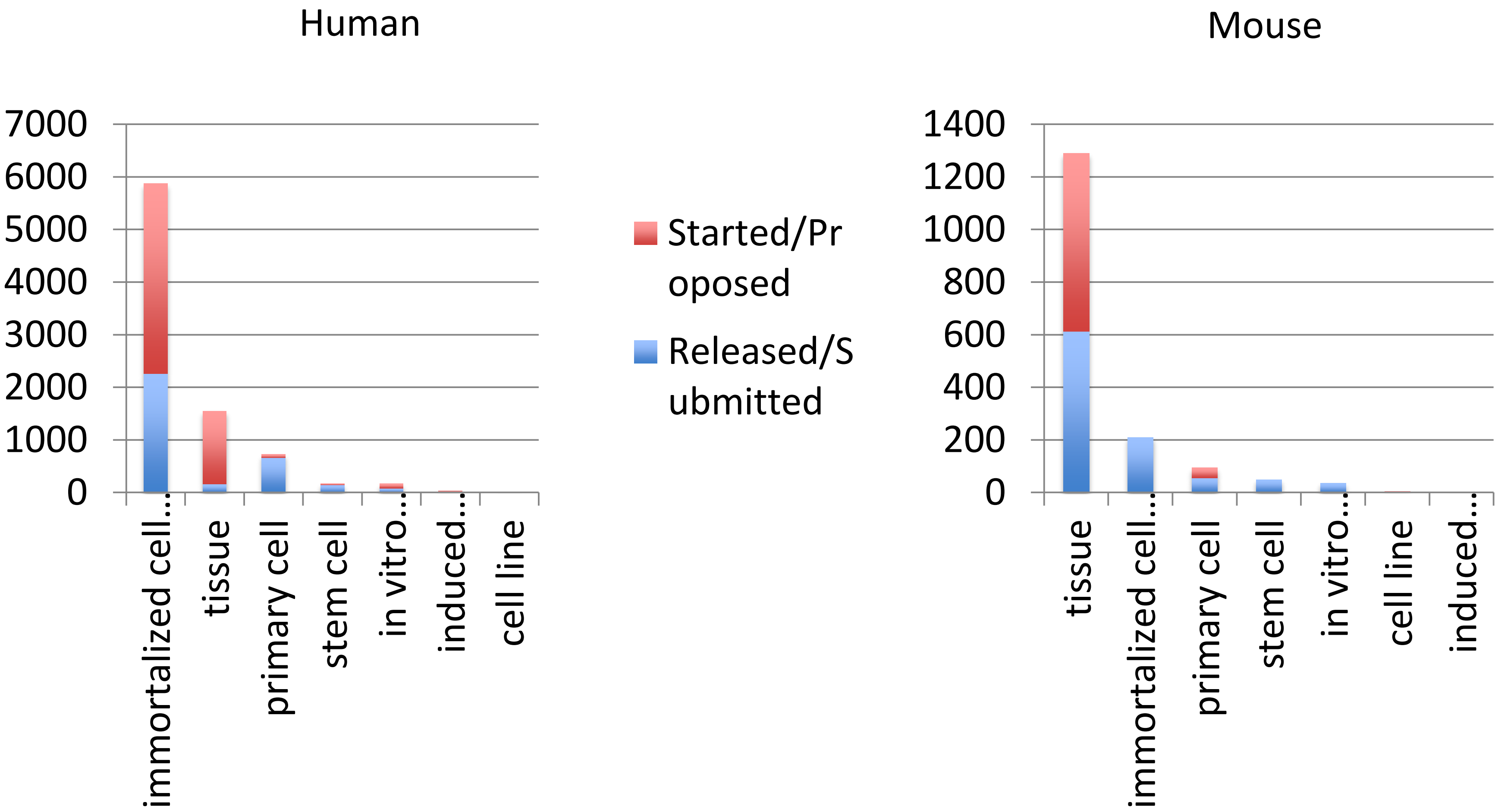


Comparing Mouse and Human with Mouse ENCODE Data

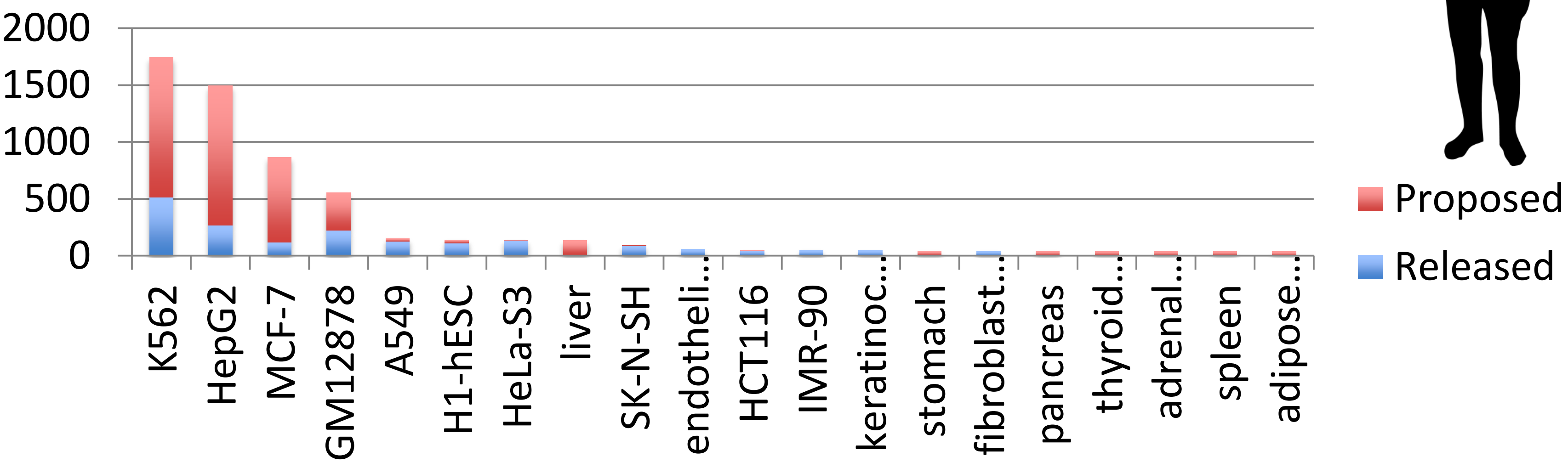


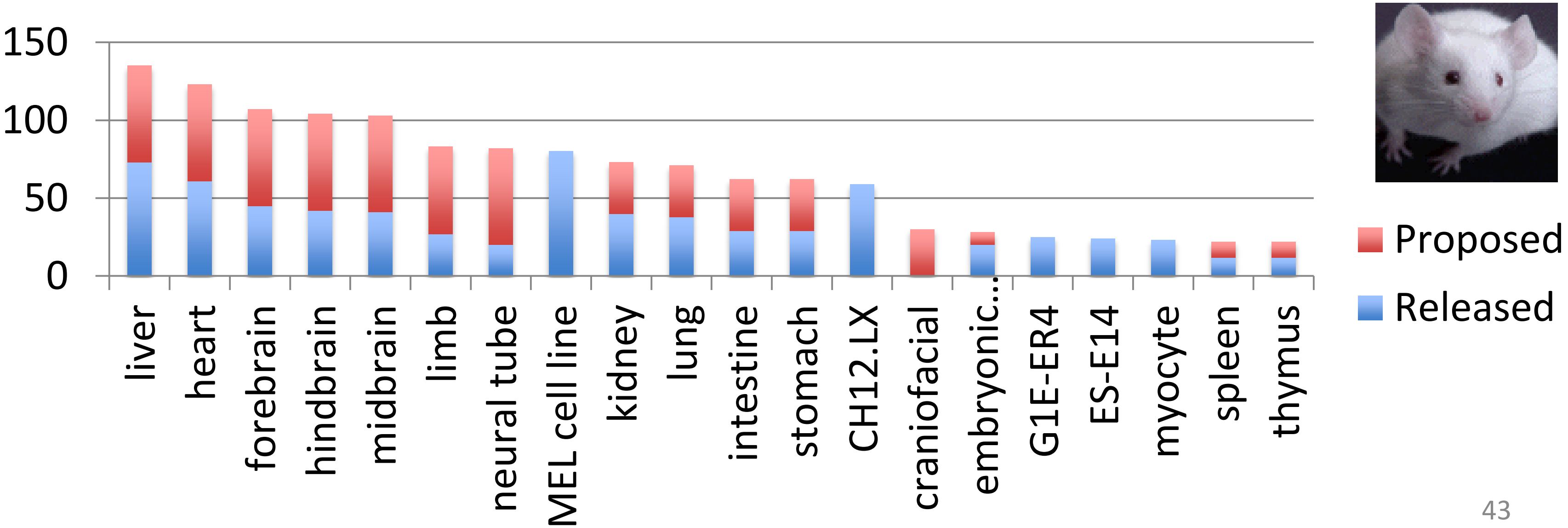


Number of Datasets Per Biosamples



Assays Per Biosample


K562
HepG2
MCF-7
GM12878
A549
H1-hESC
HeLa-S3
liver
SK-N-SH
endotheli...
HCT116
IMR-90
keratinoc...
stomach
fibroblast...
pancreas
thyroid...
adrenal...
spleen
adipose...
Proposed
Released


liver
heart
forebrain
hindbrain
midbrain
limb
neural tube
MEL cell line
kidney
lung
intestine
stomach
CH12.LX
craniofacial
embryonic...
G1E-ER4
ES-E14
myocyte
spleen
thymus
Proposed
Released